T H E S E D E D O C T O R A T D ' E T A T

ES SCIENCES MATHEMATIQUES

présentée

à l'Université Pierre et Marie Curie

- Paris 6 -

par

TAREK MOUSSA

pour obtenir le grade de DOCTEUR ès SCIENCES

Sujet de la Thèse :

L'analyse factorielle des correspondances et la reconnaissance des formes :

Analyse, codage, segmentation et reconnaissance de la parole continue.

Soutenue le 24 Mai 1982

devant le jury composé de :

M. D. DUGUE PRESIDENT
M. J.P. BENZECRI
M. J. CELEYRETTE
M. G. CHARBONNEAU
M. P. GRIVET
M. K. KREWERAS
M. X. RODET

THESE DE DOCTORAT D'ETAT

ES SCIENCES MATHEMATIQUES

présentée

à l'Université Pierre et Marie Curie

- Paris 6 -

par

TAREK MOUSSA

pour obtenir le grade de DOCTEUR ès SCIENCES

Sujet de la Thèse :

L'analyse factorielle des correspondances et la reconnaissance des formes :

Analyse, codage, segmentation et reconnaissance de la parole continue.

Soutenue le 24 Mai 1982

Devant le jury composé de :

M. D. DUGUE PRESIDENT
M.J.P. BENZECRI
M. J. CELEYRETTE
M. G. CHARBONNEAU
M. P. GRIVET
M. K. KREWERAS
M. X. RODET

J'adresse mes remerciements à :

Monsieur Le Professeur D. DUGUE pour m'avoir fait l'honneur de présider le jury de cette thèse.

Monsieur Le Professeur J.P. BENZECRI qui a dirigé ce travail et qui m'a permis, par la richesse de son enseignement, de le mener à terme.

Monsieur G. CHARBONNEAU, Maître-Assistant, qui a dirigé quotidiennement mes recherches, dont l'aide et l'amitié m'ont été très précieuses.

Messieurs Les Professeurs P. GRIVET, Membre de l'Académie des Sciences, J. CELEYRETTE, Président de l'Université de Lille III, G.KREWERAS et X. RODET pour avoir accepté de faire partie du jury.

Monsieur Le Professeur M. SAUZADE, Directeur de l'Institut d'Electronique Fondamentale, qui m'a accueilli dans son Laboratoire.

Je témoigne une reconnaissance particulière à Madame SUDRAUD et Monsieur D. SAVARY pour leur aide et leur amitié.

Je remercie Madame C. MAZET pour le soin qu'elle a apporté à la frappe de ce mémoire ainsi que les personnes qui ont effectué le tirage et la reliure de cette thèse.

Je suis reconnaissant aux autres personnes qui m'ont aidé, qu'elles veuillent bien me pardonner de ne pas les nommer ... la liste serait trop longue.

Alissar ...

mes rêves ...

INTRODUCTION

Dans l'histoire de la phonétique de la parole, le progrès des techniques permet de distinguer plusieurs étapes.

La première étape va de la découverte de l'écriture alphabétique en Phénicie, il y a plus de 30 siècles, jusqu'à la fin du XIXème siècle : c'est l'étude de la parole sans l'aide d'instrument physique. Il est vraisemblable que les premiers travaux ne furent qu'empiriques, mais on est certain que dès l'époque hellénistique, notamment à Edesse, il existait une science phonétique, dont toutefois aucun document considérable n'a été conservé. Les travaux des phonéticiens de l'Inde ancienne sont en revanche très bien connus, et leur rencontre avec la science européenne du XIXème siècle joua un rôle décisif dans la naissance de la linguistique moderne.

La deuxième étape est celle de l'enregistrement quasi-direct des vibrations de l'air et de leur observation.

Préparée par les travaux d'acoustique et de psychoacoustique du grand physicien Helmholtz , et les découvertes de la téléphonie et de la phonographie, l'étude descriptive des sons du langage parlé naît à Paris au seuil de notre siècle, des travaux de l'Abbé Rousselot puis de ceux de Gramont. Armés seulement de modestes appareils - une membrane vibrant au son commande un stylet qui dessine l'onde sur un cylindre noirci de fumée - les chercheurs de cette Ecole, bientôt répandus dans le monde entier, ont accumulé les découvertes par l'étude patiente des courbes, conjuguée avec l'observation directe par l'oreille , mais ils n'ont pas épuisé l'étude physique de la parole.

.../...

La troisième étape est celle de l'étude électronique de la chaîne parlée. Cette étude évidemment dépendante des progrès de la téléphonie, peut être symbolisée par l'usage d'un instrument : le sonographe. Mis au point aux Etats-Unis par les laboratoires Bell, le sonographe, aujourd'hui mondialement usité, fournit du signal sonore une représentation visible, comme la succession des spectres instantanés de puissance répartis en quelques dizaines de bandes de fréquence et figurés en noir et blanc, ou plutôt en gris dégradés. Fournissant directement et rapidement les mesures que l'on devait patiemment extraire des courbes de l'Abbé Rousselot, et avec elles bien des lignes nouvelles, le sonographe a notamment révélé dans la parole humaine une structure en formants, c'est-à-dire en bandes de fréquence où se concentre presque toute l'énergie, la place et les glissements de ces fréquences définissant respectivement les consonnes et les voyelles. De ce point de vue, l'étude des langues les plus usitées de l'Europe, a été menée à bien par le Professeur Delattre, savant français, établi au Etats-Unis. Mais il resterait beaucoup à faire pour mener à un même degré de perfection l'étude d'autres langues, notamment celle de l'Arabe.

La quatrième étape est celle de l'étude informatique de la parole. Il importe de rappeler ici que l'étude électronique de la parole est comprise dans ce que les techniciens nomment traitement analogique du signal, c'est-à-dire traitement de la quantité continue variable comme telle et non traduite en chiffres ; cette traduction préalable au traitement numérique n'a débuté qu'avec l'avènement des ordinateurs de grande puissance. Sur le traitement analogique, le traitement digital l'emporte par une beaucoup plus grande rigueur et une meilleure reproductibilité : un système analogique en effet, comporte généralement un système complexe de circuits dont le fonctionnement n'est qu'incomplètement analysé : quand, par exemple, un détecteur de mélodie présente sur un oscillographe la ligne d'intonation d'une phrase, on ne sait au juste quelles interpolations ont conduit de cette fonction du temps qu'est le signal sonore, à cette autre fonction que porte l'écran fluorescent.

Cependant le traitement digital a été jusqu'ici relativement peu employé ; et cela pour plusieurs raisons assez diverses. D'une part, des équipes de chercheurs encore jeunes ont été formées au traitement analogique, et en conservent les techniques. D'autre part, le traitement digital n'est en réalité

véritablement novateur que pour autant que les informations multiples et précises qu'il fournit, sont considérées comme telles et non exploitées comme le seraient les résultats issus d'une élaboration analogique. Mais pour que l'esprit humain soit instruit et non submergé, par un flux de nombres, il est impérativement requis que celui-ci soit immédiatement organisé et analysé. Or les techniques nouvelles de l'analyse statistique des données, techniques, elles aussi subordonnées à l'usage de l'ordinateur, sont récentes et, par conséquent, encore peu répandues.

Ainsi l'analyse des données doit-elle constituer une cinquième étape dans l'acoustique linguistique.

Un programme ambitieux pourrait être le suivant :

1 - Réaliser en traitement digital l'équivalent d'un sonographe. Pour la spectrographie, les programmes d'analyse de Fourier rapide, sont aisément applicables à la parole, et il est facile de réaliser un codeur digital à un nombre variable de canaux, couvrant un domaine spectral plus étendu que celui de la bande téléphonique généralement étudiée en traitement analogique.

2 - Par l'analyse des correspondances, substituer à la reconnaissance des formants quelque peu entachée de subjectivité, le calcul des facteurs dont les premières expériences ont montré la supériorité. Dans cette nouvelle schématisation du signal vocal, les définitions des voyelles et des consonnes devront acquérir une rigueur, permettant la segmentation, la décomposition automatiques de la chaîne parlée en des unités d'articulation en phonèmes. Il importe de souligner que la dictée phonétique visée ici, acceptera nécessairement les mots tels qu'ils sont prononcés, et non tels qu'on convient qu'ils devraient l'être ; ce sera l'occasion de très utiles réflexions linguistiques non dépourvues d'analogies avec celles qui présidèrent à la découverte de l'écriture phonétique.

3 - Nous parlons jusqu'ici des voyelles et des consonnes comme si la réalisation pour tout sujet était la même. Il n'en est évidemment rien. Dans le cadre d'une étude statistique de la parole digitalisée, la variabilité des prononciations devient elle-même un objet d'études rigoureuses. Il s'agit de découvrir de quels paramètes - variant avec l'individu - dépendent les axes factoriels sur lesquels nous devons projeter de nouveaux spectres instantanés afin d'en calculer les facteurs que nous substituons aux formants pour segmenter et reconnaître le nouveau signal d'après l'analyse statistique d'un bref échantillon de parole.

.../...

L'étude considérée découvrira des structures simples qu'il sera possible alors d'utiliser notamment pour coder, segmenter et reconnaître la parole.

Pour diverses considérations, nous avons choisi d'étudier la langue arabe, mais comme le lecteur pourra lui-même le constater, les méthodes décrites dans cette thèse sont indépendantes du langage utilisé ou à tout le moins facilement adaptables à une autre langue.

Nous sommes persuadés que ce travail contribue, par les méthodes mises en oeuvre, aux études scientifiques de la phonétique arabe à la segmentation et la reconnaissance de la parole continue.

La rédaction de ce mémoire axée naturellement sur les résultats positifs ou prometteurs signale également certains de nos nombreux essais infructueux et quelques projets à suivre.

Mais pour que ce mémoire ne devienne, vue la très grande taille des données traitées, un recueil de listings et de tracés, nous avons décidé de n'en publier que des extraits et le minimum nécessaire.

Cet ouvrage comporte une introduction où nous rappelons les étapes historiques de la phonétique de la parole. Dans les chapitres qui suivent, nous exposons la production de la parole naturelle, sa perception, les classifications phonétiques et un rappel des recherches sur la reconnaissance automatique de la parole.

Ensuite, nous décrivons la préparation du signal acoustique pour son traitement par l'analyse factorielle des correspondances que nous présentons rapidement.

Puis nous exposons nos travaux sur l'analyse de la parole continue, sa paramétrisation, sa segmentation et sa reconnaissance automatique.

CHAPITRE A

LA PAROLE

La communication orale résulte de trois composantes fondamentales : le signal acoustique phonétique, la structure linguistique et la perception auditive et psychologique du message sonore. Toute l'étude de la parole repose sur l'analyse de l'une ou de l'ensemble de ces trois composantes.

Nous nous limitons, dans notre étude de la parole arabe, au seul phénomène acoustique. Nous approchons ainsi, sans le vouloir -ni le prétendre- et avec des méthodes quantitatives qui ne leurs étaient pas habituelles, les phonéticiens de la phonétique traditionnelle que Cantineau (1960) définit comme "l'étude des sons concrets du langage parlé, envisagés dans leur mode de production, puis en tant que phénomènes acoustiques, et enfin dans leur audition".

Pour mieux comprendre les caractéristiques, la structure de la parole continue, nous présentons rapidement l'appareil phonatoire, le mode de production du langage et les différentes catégories de sons du langage.

A.1. L'APPAREIL PHONATOIRE

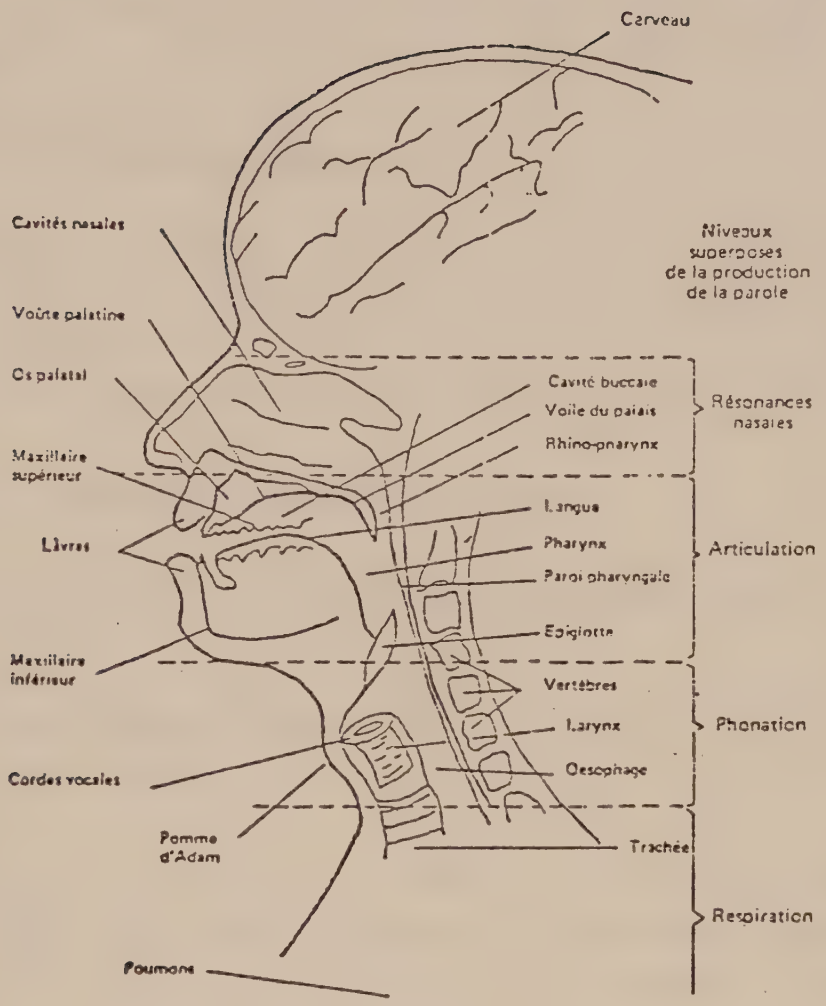
L'appareil phonatoire (Figure A.1) ressemble à un instrument à vent où nous pouvons distinguer une source d'énergie, un système d'excitation servant à mettre le flot d'air en vibration et un système de résonateurs qui modifient le spectre du signal d'excitation.

- Les poumons constituent l'élément moteur de la phonation en fournissant l'essentiel de l'énergie nécessaire sous forme d'un courant d'air expiré par la trachée-artère ; ils gèrent la pression de cet air et son débit.

- Le larynx, se situant à l'extrêmité supérieure de la trachée, est constitué d'un ensemble de muscles et de cartilages mobiles. Il contient les cordes vocales : deux lèvres symétriques qui peuvent obstruer complètement le larynx, ou au contraire en s'écartant, former une ouverture triangulaire appelée glotte. Au-dessus des cordes vocales, existe un autre rétrécissement du larynx constitué par "les fausses cordes vocales", encore appelées bandes ventriculaires qui peuvent retarder le passage de l'air donnant ainsi le temps à la glotte de prendre la configuration voulue. Au sommet du larynx se trouve l'épiglotte, un petit opercule qui se rabat pendant la déglutition, participant à la fermeture du larynx.

- Le conduit vocal, ou le corps sonore de l'appareil phonatoire, est une cavité composite s'étendant depuis les cordes vocales jusqu'aux lèvres avec en parallèle les fosses nasales ; il comprend le pharynx situé entre la racine de la langue et les parois de l'oesophage. Les cavités nasales peuvent être ouvertes ou fermées suivant la position du voile du palais, membrane que nous voyons au fond de la bouche dont l'extrêmité inférieure est la luette. Le conduit buccal comprend le palais, la langue, les dents et les lèvres. La forme et le volume des cavités buccales sont commandés par l'écartement des mâchoires, la position et la forme de la langue, la position et la forme des lèvres. Ajoutons que la luette, la pointe de la langue et les lèvres peuvent entrer en vibration.

.../...



Coupe médiane schématique des organes de la parole.

Figure A-1

(d'après Baudry M.)

A.2. LA PHONATION

Parler, articuler un mot, faire fonctionner les éléments de l'appareil phonatoire, c'est réaliser instantanément un programme très élaboré de mouvements qui déterminent les variations de forme, de dimension et de position de ces éléments qu'ils soient actifs ou passifs dans la production des signaux acoustiques de la parole.

A.2.1. L'appareil respiratoire

Pendant la phonation, les poumons exercent sur l'air à expirer une pression variable avec la force exercée par la cage thoracique et avec l'ouverture de la glotte. L'appareil respiratoire joue, aussi, un rôle dans le rythme syllabique et dans celui de la phrase ainsi que dans l'intonation (Straka G.) et (Collier R.).

A.2.2. Les cordes vocales et la glotte

Les cordes vocales agissent comme un modulateur d'air.

Au cours de la respiration normale et la voix chuchotée, elles sont largement écartées, l'air y passe librement ; de même, durant la phonation, des sons dits "sourds" ou "non voisés".

Au contraire, au cours de la phonation des sons dits "voisés", les cordes vocales vibrent à une fréquence dépendant de leur tension. Lorsqu'elles sont faiblement tendues, la fermeture est en général incomplète et l'amplitude assez grande, ce qui entraîne une fréquence grave. Mais lorsqu'elles sont fortement tendues, la fréquence est élevée et seuls les bords vibrent.

La fréquence de vibration F_o des cordes vocales, dite fréquence fondamentale, détermine principalement la hauteur de la voix. Les variations de cette hauteur constitue la mélodie du langage parlé. La fréquence fondamentale varie pour un homme adulte entre 80 et 200 Hz et pour une femme entre 160 et 400 Hz. La plage de variation dépend à la fois de l'individu et aussi de la façon de parler.

La pression sous-glottique, la fermeture plus ou moins complète des cordes vocales, responsables du voisement, et l'amplitude de leur mouvement font varier l'intensité du son produit.

.../...

A.2.3. Le conduit vocal

A.2.3.1. Excitation du conduit vocal

Il existe d'autres types de sources sonores.

L'excitation peut être causée par un flot turbulent d'air en certains points de resserrement (constrictions) du conduit vocal. Le signal d'excitation n'est pas périodique : il s'agit d'un bruit. Les constrictions peuvent se produire en divers endroits du conduit vocal et sont caractéristiques des sons produits dits fricatifs.

Une autre excitation peut être due à l'ouverture brusque d'une occlusion dans les cavités buccales qui provoque une augmentation de pression et ainsi la production d'une plosive ou occlusive.

Notons que l'un ou l'autre de ces deux derniers types d'excitations peut coexiter et se produire avec le voisement.

A.2.3.2. Transmission du conduit vocal

Au cours de l'élocution de parole continue, les propriétés de transmissions acoustiques du conduit vocal sont essentiellement dues aux variations de section du conduit vocal et à l'ouverture de la luette vers les fosses nasales.

Le spectre en fréquence de la fonction de transfert du conduit vocal présente ainsi des bosses, des maxima, des formants. Dans la parole continue, les formants varient constamment. Néanmoins, comme pour l'appareil phonatoire, il existe des termes articulatoires et acoustiques. Les sons associés à ces termes sont les éléments les plus petits discernables appelés phonèmes. Ils n'ont pas toujours une réalité bien définie. Cependant, si dans une suite de mots, nous supprimons les états stables, l'intelligibilité est très faible (Expériences de Boulanger, citée par M. Rossi dans "Table ronde sur la linguistique", Actes de la "Journée d'Etude de la Communication parlée", Bruxelles, Mai 1973 d'après X. Rodet 1977).

.../...

En revanche, si nous supprimons les transitions en maintenant les états stables, elle demeure bonne. Certains éléments sonores ont des différences perceptibles qui toutefois n'en font pas des phonèmes différents. Ils sont appelés des variantes d'un phonème.

Dans tous les cas, il est impossible de trouver un schéma unique et précis entre un phonème et sa réalisation acoustique ou articulatoire en raison du caractère continu de la parole et donc du contexte phonémique. Nous pensons que, malgré tout, le phonème est l'élément de base dans la chaîne parlée ; d'ailleurs, il est souvent possible de segmenter visuellement le signal acoustique en des phonèmes successifs.

Nous reviendrons en détail sur la question des unités phonétiques quand nous traiterons le problème de la segmentation (Chapître F).

A.3. LA CLASSIFICATION PHONETIQUE

La phonétique répartit les différents constituants du langage parlé en deux grandes classes : les consonnes et les voyelles.

Mais ce sont les modes articulatoires qui caractérisent les sons produits et permettent des classifications plus fines dans chacune des deux grandes classes. Ces modes articulatoires résultent des diverses possibilités d'expiration phonatoire. Le passage de l'air peut être libre ou non, soit à la glotte, soit au conduit vocal. D'autre part, le nombre, trois en série, un en parallèle, et les formes des résonateurs modifient l'écoulement de l'air et déterminent le son produit.

Au niveau de la glotte, si les cordes vocales vibrent, les articulations sont dites voisées ou vocaliques ou sonores (pour les consonnes). Dans le cas contraire, elles sont dites non voisées ou sourdes (pour les consonnes).

En ce qui concerne le conduit vocal, il y a trois possibilités : il est libre, c'est le cas des voyelles, il est rétréci en un ou plusieurs points, c'est le cas des articulations fricatives qui peuvent être réparties en plusieurs groupes suivant les points d'articulation, ou totalement obstrué : consonnes occlusives ou plosives. Les deux niveaux, glotte et conduit vocal, étant indépendants, tous les croisements sont possibles .

.../...

Enfin, il existe un mode d'articulation binaire : l'ouverture ou la fermeture du voile du palais donne naissance à des articulations dites nasales ou orales.

L'articulation peut être :

- *Bilabiale* : les deux lèvres s'appliquent l'une contre l'autre.
- *Labiodentale* : la lèvre inférieure s'applique sur les incisives supérieures.
- *Interdentale* : les incisives sont légèrement écartées et la pointe de la langue passe entre elles.
- *Dentale* : la pointe de la langue s'applique sur les incisives supérieures ou leurs alvéoles.
- *Pré-palatale* : la langue vient en contact avec la partie antérieure du palais.
- *Post-palatale* : le dos de la langue s'applique sur la partie postérieure du palais.
- *Vélaire* : le dos de la langue s'applique sur le voile du palais et la luette.
- *Pharyngale* : le pharynx se rétrécit par constriction de ses parois.
- *Laryngale ou glottale* : le larynx ou plutôt la glotte se ferme ou s'ouvre plus ou moins.
- *Latérale* : l'air passe sur les côtés de la langue.

A.3.1. Les consonnes

L'arabe littéraire et non le dialectal possède un système de 28 consonnes primitives selon le tableau (A-2) mais certains phonéticiens, tenant compte des variantes possibles qui dépendent du contexte articulatoire, de la locution et de la phonation incorrecte, arrivent à un total de 48 consonnes et même à 50 (Cantineau J.).

Généralement, les altérations phonétiques possibles qu'un phonème pourrait subir, sont inconditionnées et peuvent se produire dans n'importe quel contexte phonémique, ou conditionnées donc produites par le voisinage de certains phonèmes.

.../...

Bien que nous nous intéressions à l'arabe non dialectal, il est pratiquement impossible de trouver une phonétique arabe qui ne soit pas altérée soit par des caractéristiques régionales, soit même par des modes de phonations comme l'emphatisation ou autre qui peuvent altérer des consonnes aussi bien que des voyelles.

Précisons que la description des consonnes donnée par le tableau A2 ne requiert pas l'unanimité des phonéticiens arabes et qu'un important travail reste à faire dans ce domaine descriptif de la phonation arabe.

A.3.2. Les voyelles

Pour les voyelles, seule la forme du conduit vocal peut les distinguer puisque la source sonore est toujours la glotte, sauf en voix chuchotée. Cette forme conditionne les formants du spectre dont les positions qui dépendent de la fréquence fondamentale, du locuteur et du contexte, caractérisent les voyelles. Entrent principalement en compte dans les classifications des voyelles, les formes du conduit vocal : la position de la langue, la forme des lèvres, l'ouverture du conduit et la position du voile du palais ; s'ajoute à ces critères la tenue ou la durée de la voyelle.

L'arabe ne connaît que trois sons vocaliques longs ou brefs : *a* , *i* , *u* . Le son *a* l'emporte de beaucoup ; des statistiques lui donnent de 55 à 60%, *i* et *u* viennent ensuite à égalité (Fleisch H. 1956). Massignon (Communication verbale citée par Monteil V. 1960) donne des voyelles brèves des statistiques assez proches.

Mais, à ces voyelles "cardinales" s'ajoutent d'autres voyelles dues à des phénomènes combinatoires : sous l'influence des consonnes voisines, une voyelle peut subir diverses altérations et les plus exposées à ces altérations, sont les voyelles longues que certains assimileraient à des diphtongues formées par exemple, dans le cas du *u* long, de la voyelle courte *u* et de la semi voyelle *w* et dans le cas du *i* long, de la voyelle courte *i* et de la semi voyelle *y* . A remarquer que des voyelles brèves sont devenues , parfois, ultra-brèves. En revanche, c'est le *a* et surtout le *a* long qui est le plus exposé aux variations ; nous trouvons ainsi des variantes de *a* entre le *a* et le *é*, même entre *a* et *en*. Le *i* long peut parfois s'approcher du *u* long (Cantineau J.).
Il faut ajouter que certains dialectologues ont pu distinguer 17 et même 20 sons vocaliques (Marçais W.).

.../...

A.3.3. Les semi-voyelles

Comme dans bien d'autres langages, l'arabe possède deux semi-voyelles que nous avons présentées au tableau des consonnes.

A.3.4. Les diphtongues

L'arabe possède deux diphtongues fondamentales, à premier élément bref : le *ay* et *aw* mais il arrive fréquemment que leur phonation soit partielle ; le premier élément *a* s'adapte au second donc en *e* devant *y* et *o* devant *w* ; le second élément tendant à devenir extrêmement bref jusqu'à même tomber dans certaines phonétiques arabes.

Exemples : moz au lieu de mawz et mot au lieu de mawt.

Alphabet arabe	Notation machine	occlusive	fricative	nasale	liquide	voisée ou sonore	non voisée ou sourde	emphatique	semi-voyelle	bilabiale	labio-dentale	interdentale	dentale	prépalatale	postpalatale	vélaire	pharyngale	glottale ou laryngale
ء	E	×					×											×
ب	B	×				×				×								
ت	T	×					×						×					
ث	C		×				×					×						
ج	J		×			×								×				
ح	>		×				×										×	
خ	<		×				×									×		
د	D	×				×							×					
ذ	;		×			×						×						
ر	R				×	×								×				
ز	Z		×			×							×					
س	S		×				×						×					
ش	X		×				×							×				
ص	$		×				×	×					×					
ض	&		×			×		×						×				
ط	=	×					×	×					×					
ظ	≠		×			×	×	×				×						
ع	)		×			×											×	
غ	(		×			×										×		
ف	F		×				×				×							
ق	Q	×				×	×									×		
ك	K	×					×							×	×			
ل	L				×	×								×				
م	M			×		×				×								
ن	N			×		×							×					
ه	H		×				×											×
و	W					×			×	×								
ي	Y					×			×					×				

Tableau A-2

Description des sons arabes

(d'après Cantineau J.)

A.4. AUDITION ET PERCEPTION DE LA PAROLE

En parole continue, les phonèmes s'enchaînent rapidement et la distinction entre termes et transitions n'est pas toujours claire. Il en résulte parfois un chevauchement qui augmente l'influence d'un phonème sur ses voisins. Bien que le phonème ait une réalité perceptuelle et psychologique certaine, ce n'est pas exclusivement en tant que chaîne de phonèmes que la parole est perçue.

L'intelligibilité du message parlé est assurée par l'auditeur grâce probablement à une recomposition de ce message impliquant des données linguistiques et relatives à l'éducation de l'individu (sa culture), grâce aussi à la grande redondance qui caractérise la parole.

Bien que les mécanismes de fonctionnement de l'oreille et la nature exacte des informations transmises au cerveau ainsi que le traitement de ces informations soient loin d'être parfaitement connus, la perception de la communication parlée passe principalement par deux niveaux.

Le niveau acoustique n'étant que la première étape, l'auditeur fait, au besoin une analyse lexicale, syntaxique ou sémantique pour recomposer le message et le comprendre. Il peut lever l'ambiguité acoustique due à un bruit parasite, à des déformations, défauts de prononciation, perturbations ou insuffisances acoustiques, par le recours à des informations linguistiques. Il ne faut pas oublier l'information apportée par la mélodie ou l'intonation, la variation du débit et parfois d'autres informations visuelles comme les mouvements des lèvres ou les attitudes qui sont des éléments non négligeables dans la compréhension du message parlé.

Il semble que dans la perception de la communication parlée, plusieurs mécanismes différents puissent jouer à la fois ou en fonction des situations pour inférer une décision finale à partir d'indices distordus mais surabondants. Il est donc difficile de réaliser un système de reconnaissance automatique de la parole analogue à celui de l'homme. En effet, un calculateur ne permet guère de simuler fidèlement les opérations nerveuses et mentales d'êtres vivants où un grand nombre d'informations sont traitées en parallèle pour exploiter au mieux leur redondance et où des interlocuteurs peuvent se comprendre malgré les différences physiques entre les signaux qu'ils émettent pour prononcer le même phonème, le même mot ou la même phrase.

CHAPITRE B

LES RECHERCHES SUR LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

La parole est le moyen le plus naturel et le plus spontané d'expression et de communication dont dispose l'homme. Ce moyen est exceptionnel tant par sa vitesse, sa souplesse, que sa précision. La parole serait sans doute -et sera peut être un jour- le moyen privilégié du dialogue homme - machine, si sa complexité ne soulevait pas des problèmes considérables quant à sa reconnaissance automatique.

Pourrons-nous parler à une machine qui nous comprendrait et exécuterait nos ordres ?. La reconnaissance de la parole utilisée dans l'interrogation de fichiers, le tri postal, le contrôle de processus industriels, les langages de programmation vocale, l'enseignement assisté, particulièrement celui des langues, l'aide aux handicapés, etc ... et tant d'autres utilisations militaires et surtout civiles, conduirait peut être à un changement important d'une civilisation ou tout au moins dans la relation homme - machine. Bien plus, un système de reconnaissance de la parole constitue un schéma de base applicable à de nombreux autres signaux, cris d'animaux etc ... .

Nous trouvons dans un rapport de Carré, Haton et Liénard (1979) une liste quasi-exhaustive des applications envisageables.

Les nombreuses applications possibles dans différents domaines ainsi que le développement de l'informatique justifient cependant l'espoir de succès concernant les recherches entreprises en reconnaissance automatique de la parole.

Cette reconnaissance est liée au problème général de la reconnaissance des formes. Mais la parole, au-delà des formes sonores acoustiques qui la composent, est un système bien complexe de règles sémantiques,syntaxiques et psychologiques qui conditionnent sa perception par l'homme.

.../...

La reconnaissance automatique de la parole consiste à faire identifier, comprendre par une machine, des sons, des mots, des phrases ou un discours prononcés par un locuteur. La perception du message peut déclencher l'exécution d'un acte intelligent. L'évolution des systèmes de reconnaissance aura atteint son stade ultime lorsqu'un homme et une machine pourront dialoguer.

B.1. APPROCHES DE LA RECONNAISSANCE DE LA PAROLE

Deux approches sont généralement admises pour aborder les problèmes de la reconnaissance de la parole : l'une plus globale, l'autre plus analytique.

B.1.1. Approche globale

Dans l'approche dite globale, et qui a jusqu'à présent reçu le maximum d'attention et abouti avec plus de réussite, l'unité de base sera le plus souvent le mot, parfois la phrase, considéré comme un tout, une entité globale non décomposée.

Cette approche concerne principalement la reconnaissance des mots isolés mais peut s'étendre dans certains cas à la reconnaissance de mots prononcés en continu (par exemple une reconnaissance de nombres). Les unités ou les formes à reconnaître sont alors identifiées globalement par comparaison séquentielle aux formes prototypes d'un lexique, que la machine a préalablement apprises.

Le principal avantage de cette méthode est d'éviter les effets de co-articulation, c'est-à-dire de l'influence réciproque des sons à l'intérieur des mots. Cette méthode est la plus simple mais elle est limitée car elle nécessite l'apprentissage et donc le stockage en mémoire de tous les mots du vocabulaire. De plus, son utilisation pour reconnaître la parole continue paraît impensable. Cependant, la plupart des systèmes de reconnaissance actuellement commercialisés sont de ce type.

B.1.2. Approche analytique

Cette approche exploite la structure linguistique du message parlé ; elle tente de détecter, segmenter ce message en unités minimales : phonèmes, diphones ou syllabes et d'identifier ces composantes élémentaires qui seront considérées comme les unités de base à reconnaître.

Les deux approches ne diffèrent donc que par la définition des unités de base de la parole et aussi par le problème de la segmentation qui se pose ici et non là.

.../...

La méthode analytique ne nécessite que le stockage des composantes minimales de la langue, toujours en nombre restreint (quelques dizaines de phonèmes, quelques milliers de diphones ou de syllabes). Elle est donc plus générale et permet seule de reconnaître la parole continue. D'autre part, son adaptation au locuteur est plus facile et la représentation des mots rend la méthode moins dépendante du locuteur, à condition que les propriétés propres et invariantes des éléments de base puissent être bien caractérisées. Mais elle est beaucoup plus délicate à mettre en oeuvre à cause de sa sensibilité aux effets de coarticulation et donc aux problèmes de segmentation du message parlé en ses éléments constituants.

B.2. RAPPEL HISTORIQUE

Les premiers travaux significatifs sur la reconnaissance automatique de la parole remontent à environ 40 ans. En effet, en 1943 paraissait un article de L.L. Mjasnikov sur la "reconnaissance objective des sons de parole". Une étape importante a été franchie avec la construction du sonographe qui a fourni une représentation particulièrement féconde du signal vocal (Koenig W., Dunn H.K. et Lacey L. 1946).

Par la suite, toutes les études et les réalisations furent conditionnées par les progrès de l'électronique, puis par le développement de l'informatique et l'appel à de nombreuses disciplines et techniques diverses.

Les études sur la reconnaissance de la parole ont connu au départ une grande vague d'intérêt de la part de nombreux centres de recherches en France, en Grande-Bretagne et aux Etats-Unis puis au Japon et en Europe Continentale. Mais le problème s'est rapidement révélé beaucoup plus complexe qu'on ne pouvait le supposer initialement, les méthodes d'études ainsi que les moyens de traitements apparaissant insuffisants ou mal adaptés. Beaucoup de chercheurs se sont découragés, certains ont abandonné, d'autres se sont tournés vers des études de type linguistique ou phonétique, dont le caractère fondamental est plus affirmé.

Mais après avoir connu une éclipse au cours des années 60, au moment où les calculateurs numériques entraient dans les habitudes, et où la reconnaissance des formes recevait un fondement mathématique, les recherches ont repris un essor très net et dans de nombreux pays, en particulier aux Etats-Unis, au Japon, en Union Soviétique et puis en France. Ce regain d'intérêt a sans doute été dû

.../...

d'une part à l'avancement des recherches dans le domaine de l'analyse du son et de la parole, d'autre part, à la prise de conscience de la finalité réelle d'un tel moyen de communication. Un autre facteur de ce regain d'intérêt est la disparité croissante entre la puissance des ordinateurs actuels et les moyens dont dispose l'homme pour communiquer avec eux. Il ne fait pas de doute que la parole, canal de prédilection de l'homme pour une communication spontanée et naturelle, malgré une forte redondance interne tant acoustique que sémantique, reste par son débit informationnel, le véhicule préférentiel de la communication, sa rapidité ne pouvant être atteinte par d'autres voies qu'à la suite d'un apprentissage laborieux.

Nous ne faisons pas l'inventaire de l'évolution de la reconnaissance de la parole mais nous mentionnons les travaux marquants. On trouvera dans un article de De Mori (1978) une bibliographie détaillée et dans la thèse de Quinton P. (1980) une description des étapes les plus importantes de l'évolution de la reconnaissance de la parole.

Trois étapes sont à distinguer : les débuts jusqu'à 1960, la reconnaissance des mots isolés depuis 1960 puis la relance à partir du début des années 70 qui s'est poursuivie jusqu'à nos jours en mots isolés et surtout en parole continue.

B.2.1. Les débuts

Les premières réalisations de reconnaissance automatique de la parole remontent à 1950, date à laquelle J. Dreyfus-Graf présente son "phonétographe", un appareil d'analyse de la parole à l'aide d'un banc de filtres. A peu près à la même époque apparaissent aux Etats-Unis, des machines de reconnaissance d'un vocabulaire très limité de chiffres ou de phonèmes. Nous citons celle de Davis (1952) qui permet la reconnaissance des dix chiffres et celle de Denes (1959) qui, sur un vocabulaire de 14 phonèmes et en utilisant pour la première fois des données de type linguistique parvient à un taux de reconnaissance de 72%.

B.2.2. La reconnaissance des mots isolés

La reconnaissance des mots isolés d'un vocabulaire limité susceptible de conduire à des résultats utilisables, devenait l'objectif des chercheurs au début des années 60. Depuis, de nombreuses réalisations ont été faites. Le principe commun, avec des programmes différents, était de reconnaître des mots prononcés isolément dans un vocabulaire atteignant parfois quelques centaines de mots.

Deux contraintes majeures s'imposaient : la limitation de la taille du vocabulaire utilisé et l'indispensable adaptation préalable de la machine au locuteur. L'optimum de reconnaissance, pour un système donné, était une fonction rapidement décroissante avec la taille du vocabulaire utilisé, et le nombre de locuteurs considérés. Pour un seul locuteur, le système est capable de reconnaître en temps réel un vocabulaire de plusieurs dizaines de mots, la probabilité de bonne réponse oscillant entre 0,80 et 0,95 (Querré M., Mercier G. Gresser J.Y. 1974). Lorsque le vocabuliare augmente, le temps de réponse du système augmente -difficulté surmontable avec des calculateurs suffisamment rapides- et les performances diminuent. Et quand il s'agissait d'un locuteur unique, il était difficile de préciser lequel des deux de l'homme ou du système s'était le plus adapté à l'autre.

Signalons qu'actuellement et pour certaines applications, plusieurs machines commercialisées sont dispensées de la préalable adaptation au locuteur.

Parmi les systèmes de reconnaissance des mots isolés, citons celui de Vicens (1969), présenté à Stanford aux Etats-Unis, qui réussit un taux de reconnaissance de 92% sur un vocabulaire de 561 mots ou courtes phrases, pour un locuteur. Ce système a permis la réalisation d'un calculateur de bureau commandé par la voix et du robot manipulateur : Hand - Eye - Ear.

A Grenoble, en France, le système Tubach (1970) permet, avec l'aide de certaines règles syntaxique et sémantique, la reconnaissance presque sans erreur d'un programme Algol 60 énoncé mot par mot.

Nous pouvons citer parmi les réalisations récentes, le système Dynamo du C.N.E.T. de Lannion en France.

B.2.3. La reconnaissance de la parole continue

Si les premières recherches s'orientaient vers la reconnaissance d'un mot dans un vocabulaire limité, il était, pourtant, inévitable que les chercheurs abordent tôt ou tard le problème tout entier de la reconnaissance de la parole continue.

.../...

Au départ, l'idée directrice de la plupart des chercheurs était celle d'un codage alphabétique de la parole. Celle-ci était considérée comme une succession d'éléments adjacents bien caractérisés qui étaient souvent identifiés aux phonèmes. Mais ces derniers, symboles hérités d'autres disciplines, correspondent parfois assez mal aux objets réels parfois apparentés aux diphonèmes, ou à d'autres symboles plus complexes.

L'essentiel des efforts avait donc porté sur la segmentation de la parole et l'analyse objective des segments ainsi définis.

Ainsi, en 1965, après Sakay T. et Doshita S. (1963), Reddy (1973) avait entrepris d'extraire les phonèmes d'un flot continu de parole.

Les résultats obtenus en psychologie de la perception avaient fait ressortir l'insuffisance du codage à un niveau, de type alphabétique. Fallait-il alors pour la reconnaissance de la parole, imaginer une hiérarchie d'étapes successives allant de la réception du signal acoustique à la perception de sa signification en passant par des niveaux de phonologie, de morphologie, de syntaxe et de sémantique. L'intéraction entre ces différents niveaux traduit la redondance du signal de parole. Une redondance dont l'utilisation permet d'obtenir pour chaque niveau d'analyse, des performances supérieures à ce qu'elles seraient si ce niveau était considéré seul indépendamment des autres. Mais l'exploitation de cette redondance à tous les niveaux paraissait difficile.

Devant l'ampleur des difficultés, et la pauvreté des résultats, certains chercheurs abandonnaient la réalisation d'un système de reconnaissance pour entreprendre des études fondamentales sur la production, la perception et l'analyse objective de la parole. Gresser J.Y. et Mercier G. (1971) ont cité Stevens K.N., Zwicker E. et Oppenheim A.V. .

Nous pouvions alors distinguer deux approches : l'une mettant l'accent sur les niveaux acoustiques et phonologiques, c'est-à-dire sur le traitement du signal vocal, l'autre sur les niveaux ultérieurs. Mais les deux approches devaient être complémentaires, car une solution globale et optimale de la reconnaissance automatique de la parole ne peut se faire au détriment d'une de ces parties : traitement du signal ou analyse linguistique.

.../...

Conscients de cette réalité, en faisant porter dans le cadre de cette thèse, tous nos travaux sur le niveau acoustique, nous ne prétendons pas trouver la solution globale et optimale de la reconnaissance de la parole mais plutôt contribuer, avec des moyens simples, aux efforts entrepris dans ce domaine, dans la segmentation automatique en général et l'analyse de la parole arabe en particulier.

Les tous premiers travaux sur la parole continue et globalement sur sa reconnaissance automatique, avaient connu jusqu'en 1968 aux Etats-Unis des échecs sérieux ce qui conduisait à un certain pessimisme et surtout à des critiques dont le paroxysme fut manifestement un célèbre rapport de Pierce (1969).

Mais en 1971, un groupe présidé par Newell A. avait réagi et publié un rapport sur les "systèmes de compréhension de la parole". Ce rapport était à l'origine d'une action concertée qui a relancé et marqué le renouveau des recherches.

Par la suite, l'Advance Research Projects Agency (A.R.P.A.), organisme du Département de la Défense, finançait trois projets : celui de Bolt Beranek and Newman (BBN 1976) à Boston, celui de System Developpment Corporation (SDC) à Los Angeles (Ritea, 1974) et celui de l'Université Carnegie Mellon (CMU) à Pittsburg (Lesser, 1975).

L'objectif était l'élaboration d'un système de reconnaissance de phrases formées de mots pris dans un vocabulaire limité à 1000 mots, sans contrainte d'élocution, fonctionnant pour plusieurs locuteurs avec un taux d'erreurs inférieur à 10%.

Pour les trois groupes de recherches, il était admis qu'un système de reconnaissance performant devait traiter le signal de la parole sous tous ses aspects : acoustique, lexical, syntaxique, prosodique et sémantique.

Ainsi CMU réalisait deux systèmes : HEARSAY II (Lesser 1975, Erman 1977) et HARPY (Lawrence 1976) qui atteignaient sur le vocabulaire de 1000 mots, un taux de reconnaissance de phrases supérieur à 90%.

.../...

BBN présentait le système HWIM atteignant 45 à 50% de reconnaissance sur un langage de 1000 mots mais avec une syntaxe moins contraignante.

Enfin SDC élaborait des programmes de reconnaissance de phonèmes à 70% de réussite (Klatt, 1976).

A la même époque, les recherches se poursuivent en Europe et au Japon.

En Union Soviétique, l'évolution des recherches semble avoir été parallèle, avec moins d'à-coups peut-être ,plusieurs équipes s'y intéressent, et les travaux de reconnaissance phonétique sans contraintes linguistiques sont bien avancés ; voir le rapport de Haton (1975) et celui de Vintsink (1976).

Au Japon, plusieurs systèmes étaient réalisés ; nous citons celui de l'Université Yamanashi (Shigenaga 1978) et celui de Nippon Telegraph and Telephone Public Corporation (Shikano, 1978).

En Europe, les recherches se poursuivent en Allemagne Fédérale, en Grande-Bretange (Bridle, 1977), en Italie (De Mori 1976) et en France, où plusieurs organismes et universités s'intéressent et développent des méthodes de reconnaissance de la parole ; voir le rapport de Carré R., Haton J.P. et Liénard J.S. "Reconnaissance et synthèse de la parole. Etat de la recherche et du développement" (1979).

B.2.4. Les principaux systèmes de reconnaissance

Nous pouvons citer :

Les systèmes américains :

- Hearsay I
- Hearsay II
- Dragon
- Harpy
- Laprs
- Le système de SRI
- Le système d'IBM
- HWIM
- Le système de SDC
- Le système du Laboratoire Lincoln

Les systèmes japonais :

- Le système de l'Université Yamanashi
- Le système N T T

Le système italien de l'Université de Turin

Les systèmes français

- Myrtille (I et II)
- Esope 0
- Le système de l'ENSERG
- Le système KEAL
- Le système DYNAMO

Pour une présentation résumée des caractéristiques de ces systèmes et une étude comparative, voir Quinton (1980).

Mentionnons, quand même, que plusieurs appareils de reconnaissance de mots sont commercialisés. Nous citons quelques constructeurs : THRESHOLD (TTI, USA), INTERSTATE (USA), DIALOG (USA), NEC (Japon) dont les prix varient de 10.000 Dollars à 100.000 Dollars. D'autres constructeurs proposent des appareils moins performants et moins coûteux : CENTIGRAM, PHONICS (USA). La machine de N E C peut également reconnaître des mots enchaînés à condition de ne pas prononcer rapidement les mots courts. Certains appareils fonctionnent quel que soit le locuteur, donc sans apprentissage, pour un vocabulaire très réduit.

CHAPITRE C

TRAITEMENT NUMERIQUE DU SIGNAL ACOUSTIQUE POUR UNE ANALYSE FACTORIELLE DE LA PAROLE

Le message vocal (parlé ou même chanté) peut être considéré comme une fonction scalaire unique du temps : on pourra s'étonner que tant de richesses ne requièrent pas au moins une fonction vectorielle (à plusieurs composantes) ; mais la pratique quotidienne du traitement électronique de la parole suffit à convaincre les plus sceptiques : il s'agit bien d'un phénomène physique unidimensionnel.

Toutefois, ce n'est pas en tant que phénomène unidimensionnel que le message sonore est reçu : avant même d'être compris, ou seulement segmenté en mots ou en accords, en syllabes ou en notes, le message est analysé dans l'oreille interne et élaboré dans le cerveau. Ce n'est donc pas non plus en tant que phénomène unidimensionnel que le message sonore doit être traité par le statisticien (bien que certains chercheurs fondaient des espoirs sur l'examen direct de la courbe même du signal).Notre analyse, comme celle de l'oreille, commence par la décomposition spectrale.

En toute rigueur, le spectre instantané du signal, grandeur fonctionnelle que nous voulons substituer à la grandeur scalaire primaire (pression sur le tympan, ou magnétisation de la bande),n'est pas exactement défini. Mathématiquement, on sait sous des conditions peu embarrassantes, associer à une fonction $f(t)$ définie sur la droite réelle $(-\infty , +\infty)$ une transformée $F(\omega)$; mais c'est là le spectre de tout le message, non le spectre à l'instant t. Celui-ci ne peut être défini que d'après une tranche sonore assez brève comprenant l'instant t, et il est hors de doute qu'il ne pourra comprendre les basses fréquences associées à des périodes qui surpassent la durée de la tranche analysée !. De plus, tant le traitement numérique du signal que le traitement analogique (ou traitement comme grandeur physique continue par batteries de filtres, etc ...) n'ont pas un pouvoir de résolution temporelle infiniment fin : les très hautes fréquences (les va-et-vient accouplés en une µs par exemple) sont donc inaccessibles à nos calculs comme elles le sont à notre oreille.

Assujettie à des contraintes contradictoires, la traduction du message primaire unidimensionnel en un message fonctionnel -le spectre instantané fonction du temps- ne pourra que résulter d'un compromis, accepté après tatonnements !.

C.1. NUMERISATION DU SIGNAL

C.1.1. Choix du langage et du texte

Nous disions que pour diverses considérations (notre travail aurait dû être mené au Liban, notre langue maternelle étant l'arabe), nous avons choisi d'étudier la langue arabe mais il est facile de constater que les méthodes décrites et utilisées sont indépendantes du langage ou à tout le moins adaptables à une autre langue.

Il n'existe pas, à notre connaissance, confirmée par MM. Cohen et Gsell, des statistiques sur les fréquences relatives pour chaque phonème de la langue arabe non dialectale; cependant pour certains parlers, des chercheurs ont établi quelques statistiques (le parler du Caire (Tomiche, ...)).Il est donc impensable, actuellement, de trouver un texte phonémiquement équilibré dont la distribution phonémique correspond à celle de la langue arabe ; nous aurions pu mesurer la proximité du texte étudié à ce texte par un test de χ^2, procédure habituelle chez tous les statisticiens.

Nous ne nions pas l'intérêt que pourrait présenter un tel texte, cependant comme le montrent les études que nous avons faites et que nous détaillons plus loin (Chapitre E), il n'est pas d'une importance capitale.

Toutefois, l'équilibre phonémique, ainsi décrit, n'est qu'une approximation de premier ordre. Il n'est toujours pas suffisant lorsque des effets principaux se produisent aux transitions entre phonèmes : un équilibre relié à la statistique des diphonèmes serait alors nécessaire. Ajoutons que d'une part, les sons ne se présentent pas découpés mais enchaînés, d'autre part, ils ne se répètent jamais identiques à eux-mêmes. Cependant, en cherchant un texte réduit et représentatif, il faut s'assurer que cette réduction n'entraîne pas des effets de biais sur les résultats obtenus.

Le texte choisi pour servir de support à notre travail a été tiré d'un journal libanais, nous ne pensons pas qu'il présente une distribution phonémique anormale (cf. Annexe 2) ; d'ailleurs, les histogrammes de fréquences des

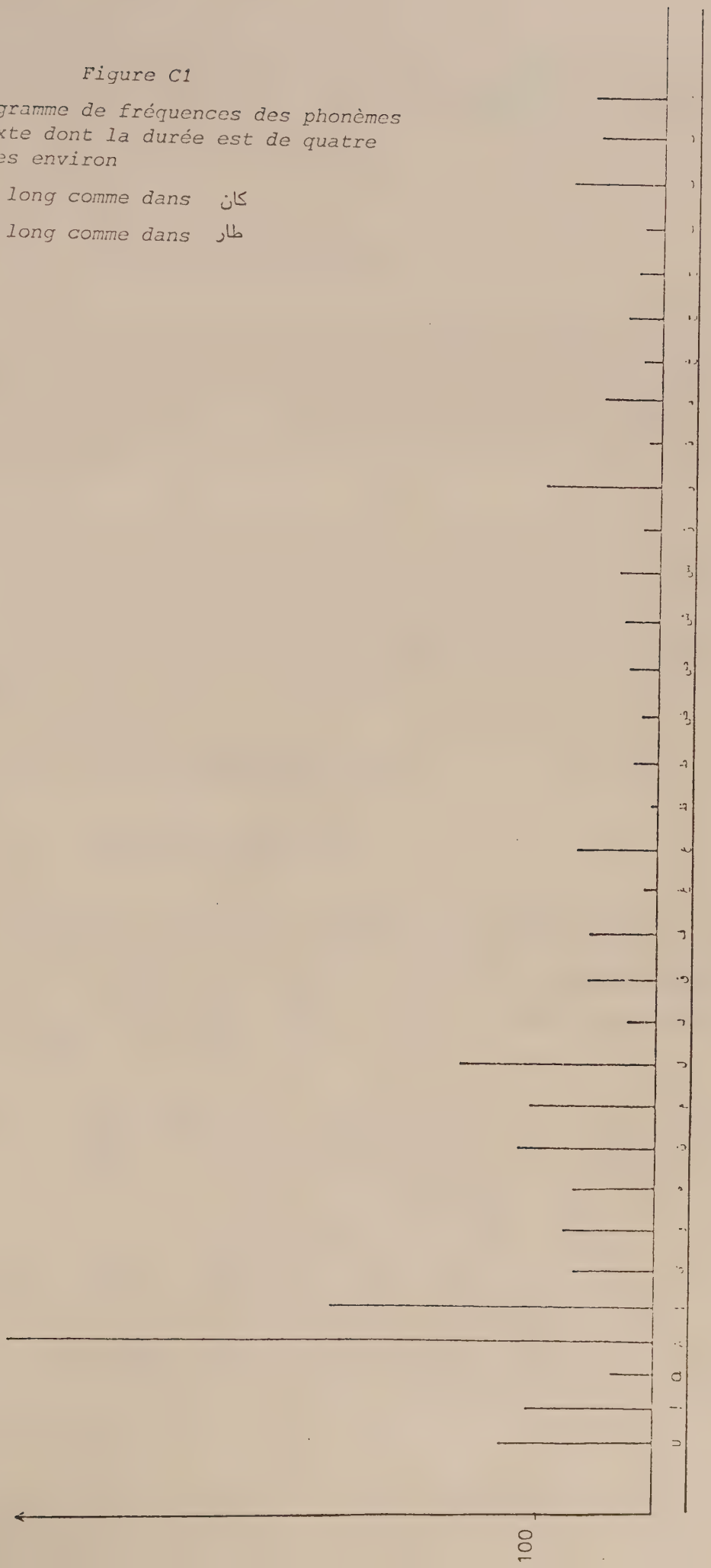
.../...

Figure C1

Histogramme de fréquences des phonèmes du texte dont la durée est de quatre minutes environ

! = e long comme dans كان

ɑ = a long comme dans طار



phonèmes, correspondant à des textes de durées successives de 25 secondes, sont pratiquement semblables entre eux et avec l'histogramme global du texte dont la durée est de 4 minutes environ *(figure C1)*.

C.1.2. Enregistrement analogique

Le texte choisi a été lu dans des conditions courantes de prononciation, c'est-à-dire avec un débit normal et une articulation naturelle, sans accentuations, ni intonations particulières ; nous discuterons plus loin les effets des changements de débit et d'accentuation dans la lecture que nous avons constatés sur le signal obtenu à partir du même texte.

Les 373 mots contenus dans l'article de presse ont été prononcés en 4 minutes, soit un débit moyen de 3 mots/2 secondes. Le texte, lu dans une salle moyennement réverbérante d'environ 50 m^3 a été enregistré à l'aide d'un microphone Beyer type M 88 N, et d'un magnétophone Revox A 77 en 19 cm / seconde (cf. Annexe 1).

C.1.3. Echantillonnage de la parole

C.1.3.1. Conditions théoriques

L'élaboration analytique des signaux ne se conçoit pas sans les notions de spectre aujourd'hui familières à tous les ingénieurs. Ces notions associées pour toujours au nom de Fourier (qui les a appliquées avec une grande généralité dans sa théorie analytique de la chaleur : 1822) se sont introduites, au milieu du XVIIIème siècle, dans les travaux d'Euler notamment.

Le traitement numérique des signaux analogiques est possible grâce à l'équivalence (Wittaker, 1915) entre la connaissance à tout instant du signal S(t) dépendant du temps, et la connaissance de ce même signal à des instants discrets S(iT), i = 1,2, ..., N où T est appelé la période d'échantillonnage. Le théorème d'échantillonnage (Shannon, 1962) précise les conditions de validité de l'équivalence : la fonction S(t) doit avoir un spectre en fréquence borné et si F_c est cette borne, la fréquence d'échantillonnage F doit être supérieure ou égale à 2 F_c. Si ces conditions ne sont pas respectées, toute fréquence f supérieure à F/2 est repliée dans la bande (0,F/2) et prend la valeur f_r telle que :

.../...

$$f_r = (n+1)\,F - f \qquad \text{si} \qquad (n + \tfrac{1}{2})\,F < f \leq (n+1)\,F \qquad n = 0,1,2, \ldots$$

$$\text{ou} \quad f_r = f - nF \qquad \text{si} \qquad nF \leq f < (n + \tfrac{1}{2})\,F \qquad n = 1,2,3, \ldots$$

Il importe donc,avant tout échantillonnage,de faire subir au signal un filtrage passe-bas éliminant aussi soigneusement que possible toute composante supérieure à la demi-fréquence d'échantillonnage choisie.

La bande passante de la parole pour l'essentiel de l'information ne dépasse guère 4000 Hz, sauf pour quelques consonnes telles que le "s" ou le "ch". Les filtres passe-bas n'ayant pas une courbe de réponse strictement rectangulaire, il est nécessaire que la fréquence de coupure du filtre soit sensiblement inférieure à la demi-fréquence d'échantillonnage.

Nous avons choisi de limiter à 7000 Hz la bande passante du signal de parole étudiée et nous avons fixé à 20480 Hz la fréquence d'échantillonnage, cette valeur étant divisible par une puissance de 2 inférieure ou égale à 4096 ce qui est souhaitable pour l'emploi de la transformée de Fourier rapide (cf. C.2.2.). Le filtre Schlumberger FAB 24 (Annexe 1) fournit une atténuation inférieure ou égale à 1 dB dans la bande passante (0 - 7000 Hz) et une atténuation supérieure à 20 dB au-delà de 10240 Hz, ce qui est suffisant pour éviter tout phénomène gênant de repli.

C.1.3.2. Conversion analogique - numérique (A/N)

Quatre minutes de parole correspondent à près de cinq millions de nombres à la fréquence d'échantillonnage que nous avons utilisée, ce qui exclut de les stocker directement dans la mémoire centrale de l'ordinateur. Il est donc nécessaire d'écrire ces nombres dans une mémoire de masse, par exemple une bande magnétique qui est un support bien adapté étant donné le caractère séquentiel de l'information et la capacité énorme du stockage (environ 50 millions d'échantillons). L'écriture de la bande magnétique nécessite cependant quelques précautions. D'une part, les échantillons sont groupés sur la bande en blocs de longueur finie et le déroulement s'arrête entre l'écriture de chaque bloc. D'autre part, les dérouleurs de bande numérique ont une vitesse de défilement insuffisamment régulière (les variations peuvent atteindre 2 à 3 %), ce qui entrainerait des distorsions dans la cadence d'échantillonnage si les nombres résultant de la conversion analogique numérique étaient envoyés directement sur la bande. De plus, il serait impossible de choisir la fréquence d'échantillonnage puisque celle-ci serait obligatoirement fixée par

.../...

le débit moyen du dérouleur de bande. Pour toutes ces raisons, il est nécessaire de découpler le rythme d'écriture des échantillons et celui de leur conversion (qui doit être rigoureusement uniforme). Ce découplage est obtenu à l'aide d'une mémoire tampon (celle de l'ordinateur) dirigée en deux tableaux. L'un des tableaux est rempli pendant que l'autre se vide grâce à deux canaux d'accès direct à la mémoire. La vitesse d'entrée des échantillons peut être alors parfaitement uniforme et pilotée par une horloge externe réglable, tandis que la vitesse de sortie vers le dérouleur est irrégulière. Les débits moyens d'entrée et de sortie sont évidemment égaux.

L'organe de conversion A/N est composé d'un échantillonneur-bloqueur Analog Devices SHA 2A et d'un convertisseur Hydrid Systems ADC 591 A de 12 bits de précision. Le temps maximal de conversion réclame au maximum 4 µs, la fréquence d'échantillonnage maximale sur le système dont nous disposons est donc de 250 KHz si les échantillons sont stockés sur un support très rapide comme un disque. Le dérouleur magnétique que nous possédons limite, toutefois, à 35 KHz la cadence d'échantillonnage.

Les signaux échantillonnés souffrent d'un bruit inhérent à la conversion A/N qui est le bruit de quantification. La conversion A/N ne fournit que des nombres entiers compris entre 0 et 4095 pour 12 bits de sorte que chaque échantillon est, à priori, entâché d'une incertitude d'arrondi inférieure ou égale à 0,5. Cette incertitude est équivalente à un bruit qui s'évalue généralement (Mathews, 1969) par comparaison avec le plus grand nombre convertible. Pour un convertisseur à 12 bits, le rapport signal sur bruit de quantification (S/B) est ainsi égal à :

$$\frac{S}{B} = \frac{4095}{0,5} \quad \text{soit 78 dB}$$

Ce bruit est comparable au niveau de bruit introduit par les meilleurs magnétophones. Il faut tenir compte cependant du fait que cette valeur correspond à un optimum obtenu avec des signaux d'amplitude maximale. Pour des signaux plus faibles, le rapport signal sur bruit diminue. Néanmoins, le signal de parole possède en langage courant une dynamique modérée (de l'ordre de 10 dB), de sorte que le bruit de quantification reste à un niveau négligeable.

.../...

C.2. ANALYSE SPECTRALE DU SIGNAL

Analyser le signal vocal est une tâche ardue tant est grand le nombre de paramètres physiques mesurables. Néanmoins, trois paramètres fondamentaux s'imposent (Moles, 1952) : l'intensité, la fréquence et le temps. Quoique non exhaustive, la représentation de l'objet sonore dans un repère formé de ces trois dimensions est féconde, car chacun des plans de référence correspond à un fait perceptif important : le plan dynamique (Intensité-Temps), le plan harmonique ou spectral (Intensité-Fréquence) et le plan mélodique (Fréquence-Temps). Nous avons utilisé un programme appelé ASPECT (Chabrel, Charbonneau, 1976) permettant les multiples traitements nécessaires à l'obtention d'éléments pouvant servir à une analyse statistique féconde de la parole.

C.2.1. Le logiciel

C.2.1.1. Structure générale du programme ASPECT

La conception du programme ASPECT a été faite avec l'objectif de satisfaire deux conditions. D'une part le programme a été écrit pour pouvoir fonctionner en mode conversationnel, c'est-à-dire qu'il soit capable d'exécuter immédiatement toute commande de l'utilisateur. D'autre part, pour que cette première condition puisse être satisfaite, il est nécessaire que les diverses directions correspondent à des transformations fournissant des éléments de même forme, de sorte qu'elles soient utilisables dans un ordre quelconque dans la mesure où celui-ci a un sens physique. ASPECT est écrit en langage FORTRAN (Dreyfus, 1969) et comporte plus de 5000 instructions découpées en 7 segments dont le programme assure, lui-même, la gestion.

C.2.1.2. Les commandes

Une commande se présente sous la forme d'un code mnémonique de deux lettres (sauf pour la transformée de Fourier inverse notée F- et la multiplication par la fonction complexe conjuguée notée M:) suivi d'une liste de paramètres séparés par des blancs ou des virgules.

L'utilisateur dispose de cinq types de commandes :

- les commandes de configuration qui permettent de fixer les valeurs de certains paramètres, comme la fréquence d'échantillonnage, les tailles des blocs, etc ...

.../...

- les commandes de traitement mathématique qui permettent d'effectuer des opérations élémentaires sur le signal.
- des commandes donnant la possibilité de créer des macro-commandes qui condensent en une seule directive un ensemble quelconque d'instructions et autorisent l'utilisation des boucles.
- des commandes de visualisation sur un écran cathodique et de tracé sur papier.
- enfin, il y a 5 macro-commandes programmées d'avance, qui assurent le calcul et la visualisation des décompositions suivantes du signal :

- Transformée de Fourier d'une partie du signal.
- Sonagramme d'une partie du signal.
- Enveloppe dynamique du son.
- Enveloppe dynamique de chaque harmonique pour les signaux harmoniques ou pseudo-périodiques.
- Recherche de la mélodie (pitch) d'une portion de la parole.

C.2.2. Le calcul des spectres de Fourier

C.2.2.1. Intérêt de l'analyse spectrale

Le traitement du message vocal nous engage à rechercher dans ses spectres des structures de deux types, structures qui sans être absolument indépendantes peuvent toutefois être considérées séparément : nous voulons parler de la mélodie ou ton (pitch des auteurs anglo-saxons)et du profil spectral.

Si la parole n'était que musique, la mélodie en serait l'air, et le profil spectral serait le timbre de l'instrument. Mais il s'agit d'un discours ; et les rôles sont presque inversés : la mélodie exprime l'intonation (accents des mots, ponctuation des phrases, etc ...) et le profil spectral caractérise le phonème ... le signal donc le texte.

D'une part, le spectre instantané présente une suite de maxima répartis en progression arithmétique ; c'est-à-dire pour parler acoustique, que ces maxima correspondent aux harmoniques sucessifs d'un fondamental ; la fréquence de celui-ci variable pour un même sujet au cours du discours et plus encore variable de sujet à sujet, de sexe à sexe, plus aigüe pour les voix féminines, est la fréquence F_o du ton.

.../...

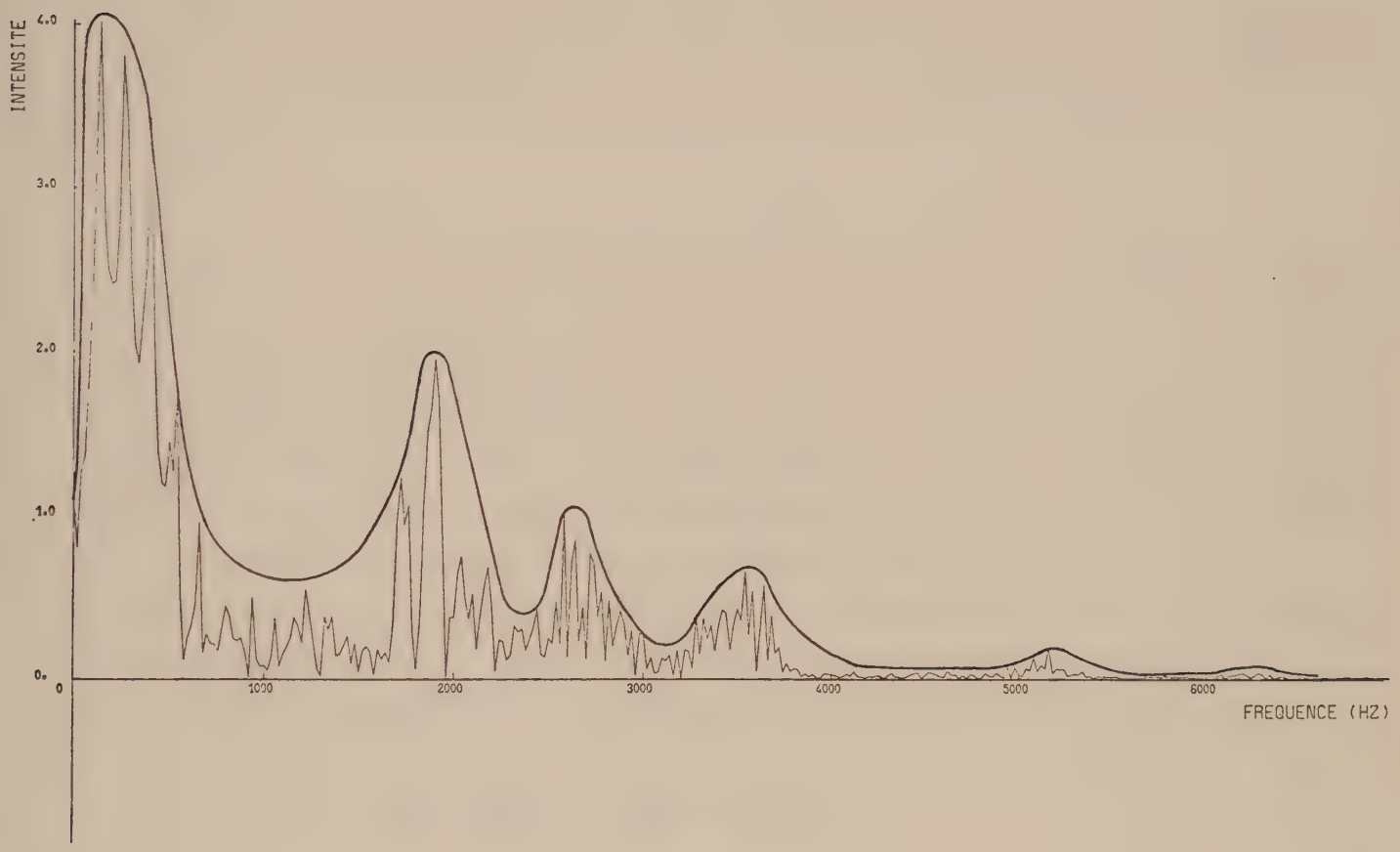


Figure C.2

Profil spectral d'une fraction d'un i
(intensité logarithmique)

D'autre part, ces maxima du spectre instantané sont de très inégales amplitudes, et viennent se loger sous une enveloppe dite profil spectral ; dont les grands maxima, larges ondulations comprenant plusieurs harmoniques du ton fondamental , sont appelés formants. Nous avons un dessin semblable à celui de la figure C-2.

Physiologiquement, le profil spectral résulte de la forme variable des résonateurs du trajet vocal et le ton est donné par la fréquence des vibrations des cordes vocales ; ces vibrations, quasi-ininterrompues et très pures dans le chant, ne sont pas constantes dans la voie parlée ordinaire ; et elles disparaissent totalement dans la voix chuchotée qui n'est qu'un bruit fricatif coloré par les résonateurs.

L'analyse spectrale est apparue, depuis longtemps déjà, comme un moyen puissant d'investigation des signaux sonores (Helmholtz, 1896). Cela est dû essentiellement au fait que les éléments résultants de cette décomposition ne sont pas des termes mathématiques arbitraires. Ils sont en relation étroite avec la perception des sons par l'oreille (Littler, 1965, p 308). Différentes études ont été réalisées (Greenwood, 1961, Plomp, 1964, ...) sur le pouvoir qu'a l'oreille de séparer une vibration complexe en ses composantes spectrales. D'une part, l'oreille se comporte comme un analyseur de spectre à large bande, peu sélectif, et d'autre part, le cerveau prolonge cette première analyse par un traitement aboutissant à une sélectivité remarquable notamment dans la détection de la hauteur (bien supérieure à la discrimination possible avec un analyseur de spectre analogique actuel).

C.2.2.2. Les formules et leurs calculs

Le théorème de la transformée de Fourier établit (Schwartz, 1963) la correspondance suivante :

$$F(\omega) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} f(t)\, e^{-j\omega t}\, dt$$

$$f(t) = \int_{-\infty}^{+\infty} F(\omega)\, e^{j\omega t}\, d\omega$$

.../...

Cette équivalence définit la représentation spectrale complexe $F(\omega)$ d'une fonction $f(t)$ sous des hypothèses très générales. La transformation ainsi définie n'est que la généralisation de la série de Fourier associée à une fonction de période T lorsque T tend vers l'infini.

En pratique, lorsqu'elle représente un phénomène physique réel, $f(t)$ n'est définie que sur un intervalle de temps fini ($0 \leqslant t \leqslant T_1$) ; de plus, si $f(t)$ est échantillonnée, l'intervalle T_1 étant divisé en K incréments égaux à $\Delta t = \frac{T_1}{K}$, on a alors une transformation de Fourier discrète telle que :

$$F(t_k) = \Delta\omega \sum_{n=0}^{n=N-1} G(\omega_n)\, e^{j\omega_n t_k}$$

avec $$G(\omega_n) = \frac{\Delta t}{2\pi} \sum_{k=0}^{k=K-1} F(t_k)\, e^{-j\omega_n t_k}$$

et $$\Delta\omega = 2\pi\,\Delta f = 2\pi\,\frac{K}{NT_1}$$

En fait, du point de vue mathématique, la transformée de Fourier discrète est un simple exercice d'algèbre linéaire, un changement de base orthonormée dans un espace de dimension finie ; toutefois, l'intérêt de ce changement de base provient de la façon dont il tend à la limite vers le développement en intégrale de Fourier lorsqu'on densifie l'échantillonnage (il est d'ailleurs de règle en mathématiques appliquées que des calculs qui ne peuvent être faits qu'en dimension finie doivent leur intérêt aux théorèmes limites relatifs aux espaces fonctionnels).

En 1965, Cooley et Tukey ont proposé un algorithme rapide pour le calcul de ces formules, le calcul étant le plus simple si N et K sont pris égaux à une puissance de 2 (en fait le principe de la méthode se trouve déjà dans Runge (1903) ; pour un historique complet de ce remarquable algorithme, cf. Cooley et Coll. (1967)). A la fin du calcul, la transformation de Fourier (commande FO) est obtenue sous la forme :

$$G(\omega_n) = a_n + j\, b_n$$

Par multiplication par la complexe conjuguée (commande M:), on obtient le spectre en puissance :

$$| G(\omega_n) |^2 = a^2_n + b^2_n$$

C.2.2.3. Choix des variables

Dans une analyse spectrale, les variables dont dispose l'utilisateur sont :

- la fréquence d'échantillonnage F.
- la fréquence maximale qui puisse être détectée F_{max}.
- la résolution en fréquence Δf.
- l'intervalle de temps Δt séparant deux échantillons.
- la durée T_1 de la fraction du signal analysée.
- le nombre de points N sur lesquels porte l'analyse (longueur d'un bloc).

Ces variables vérifient les relations suivantes :

$$F = 1 / \Delta t$$

$$F_{max} = F / 2$$

$$T_1 = N\Delta t = N / F$$

$$\Delta f = F / N = 1 / T_1$$

$$\Delta f . N . \Delta t = 1$$

Cette dernière relation est l'expression du principe d'incertitude en acoustique : il est impossible de faire une analyse fine simultanément en temps et en fréquence.

En réalité, il n'y a pas deux variables indépendantes F et N . La fréquence d'échantillonnage étant fixée à 20480 Hz pour satisfaire la bande passante du signal vocal, il ne reste de réellement variable que le nombre de points N définissant la taille des blocs servant au calcul du spectre en puissance de Fourier.

.../...

Nous nous sommes efforcés de concilier deux exigences contradictoires : d'une part, avoir une résolution en fréquence aussi bonne que possible (impliquant N grand), et d'autre part ne pas avoir des blocs d'une durée trop longue pour ne pas risquer de noyer les sons brefs (t, k, q par exemple).

Ces exigences conduisent à ne considérer que deux tailles de blocs, N = 512 qui correspond à une durée T_1 = 25 ms assurant une résolution Δf = 40 Hz, et N = 1024 correspondant à T_1 = 50 ms et Δf = 20 Hz.

C.2.2.4. Spectres successifs et spectres décalés

Pour obtenir une évolution du spectre de fréquence au cours du temps, nous avons calculé les spectres instantanés portant sur des blocs successifs de N points (1024 ou 512). Nous avons décidé également de ne pas prendre les blocs successifs mais des blocs de 1024 (ou 512) consécutifs décalés de 512 en 512 et surtout de 256 en 256 (de 256) de manière à obtenir une fenêtre sur le spectre glissant consécutivement de 12,5 ms pour suivre toute l'évolution et recueillir les meilleures informations. L'étude de ces spectres en puissance, que nous développons plus loin (chapitre E) nous a montré que nous obtenons la représentation la plus fidèle avec des spectres décalés de 256.

En effet, l'analyse d'une fraction du signal revient à considérer que le signal complet a été multiplié par une fenêtre temporelle rectangulaire valant 1 pendant le durée examinée et 0 ailleurs. Dans ces conditions, la transformée de Fourier est la convolution de la transformée de Fourier du signal par celle de la fenêtre. Les bords de la fenêtre provoquent ainsi des lobes gênants dans les spectres en fréquence (Rabiner et Gold, 1975). Pour réduire l'influence du phénomène, nous avons utilisé la fenêtre de Hanning (Kaiser, 1966) qui s'écrit :

$$h(n) = \frac{1}{2} \left(1 - \cos \frac{2\pi n}{N} \right) \qquad n = 0, 1, 2, \ldots, N$$

Si nous utilisions cette fenêtre qui annule le signal aux bords de la fenêtre, certains phénomènes brefs risquent d'être atténués fortement.
Le décalage de 256 en 256 élimine cet inconvénient car chaque phonème se trouve toujours convenablement placé dans au moins un bloc.

.../...

C.3. CODAGE DES SPECTRES POUR L'ANALYSE FACTORIELLE

C.3.1. Principe d'application de l'analyse des correspondances

Pour chacun des N points, le programme de la transformée de Fourier rapide donne N composantes :

$$\{ G(\omega_n),\ n = 1, \ldots, N \}$$

avec $\omega_n = (F / N) \times (n-1)$ Hz

comme le signal est réel, les amplitudes complexes $G(\omega_n)$ sont liées entre elles par la relation

$$G(\omega_n) = G(\omega_N - \omega_n)$$

La phase ne contribuant pas au message, seuls comptent les carrés des modules $| G(\omega_n) |^2$, ce qui réduit l'information à N/2 nombres :

$$| G(\omega_n) |^2 \quad , n = 1, 2, \ldots, N/2$$

Pour des blocs de N = 1024 points, cela fait 512 nombres à traiter d'un point de vue statistique. Cependant, les études antérieures de la parole et l'usage des vocoders à canaux indiquent qu'il est possible de réduire fortement cette quantité de variables.

Le principe de notre traitement est de découper les spectres de puissance en tranches consécutives qui sont l'analogue des canaux des codeurs usuels ; à cette différence essentielle près que nous pouvons modifier les limites de ces canaux comme des paramètres du programme de traitement, alors que dans le traitement analogique, une telle modification ne peut se faire sans changer les batteries de filtres.

Comme de règle en analyse des correspondances (cf. Chapitre D), nous constituerons un tableau I x J ; il est commode d'appeler les "i" individus et les "j" variables ; en fait ici, "i" est un spectre et "j" est un canal ; K(i,j) étant la puissance du spectre i dans le canal "j", ainsi qu'on l'explique en détail au § C.3.2. .

.../...

Les figures qui suivent présentent les spectres individuels de 6 phonèmes, ainsi qu'un spectre moyen.

Sur les figures C3a,C3b,C3c, on a les spectres de trois phonèmes consonantiques voisés ; dans ces trois spectres, le fondamental et quelques uns de ses harmoniques sont nettement visibles. On notera que l'importance relative de ces harmoniques n'est pas du tout la même ; d'où dans le profil sonore de grandes différences sur lesquelles se fonde l'analyse factorielle pour discriminer les sons. Les sons B et D objets des figures C3b et C3c ont en arabe une prononciation voisine de celle des sons français correspondants. Le son ayn (figure C3a parfois représenté dans les transcriptions par deux a consécutifs, (ex : aabd - allah) et fortement grasseyé, montre une suite remarquable de 9 harmoniques après le fondamental.

Sur les figures C3d,C3e,C3f, on a des sifflantes non voisées, caractérisées par un bruit intense dans la bande des hautes fréquences. Pour le CH (figure C3f), la bande de bruit est nettement plus basse que pour le S (figure C3e); le sad (S emphatique propre à l'arabe) se situe un peu plus bas que le S ; de plus en basse fréquence (0 - 1000 Hz), le partiel 4 est nettement visible, alors qu'il disparaît dans le S.

La figure C4, moyenne de l'ensemble des spectres, montre la répartition globale de l'énergie sonore ; c'est d'après cette figure qu'on peut en première approximation, choisir les bornes des canaux suivant lesquels on code les spectres dans l'analyse factorielle.

.../...

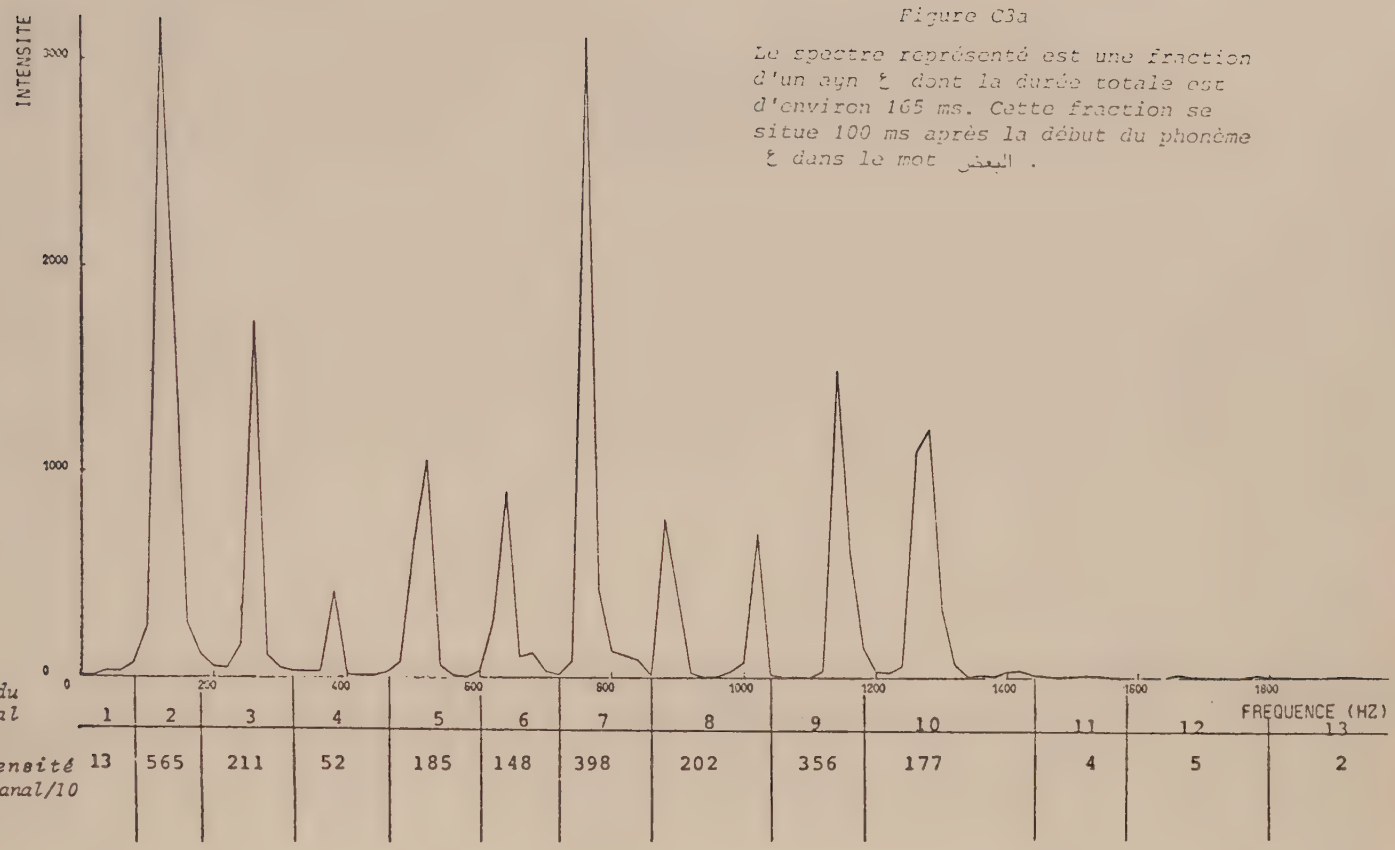


Figure C3a

Le spectre représenté est une fraction d'un ayn ع dont la durée totale est d'environ 165 ms. Cette fraction se situe 100 ms après la début du phonème ع dans le mot البعض .

N° du canal	1	2	3	4	5	6	7	8	9	10	11	12	13
Intensité du canal/10	13	565	211	52	185	148	398	202	356	177	4	5	2

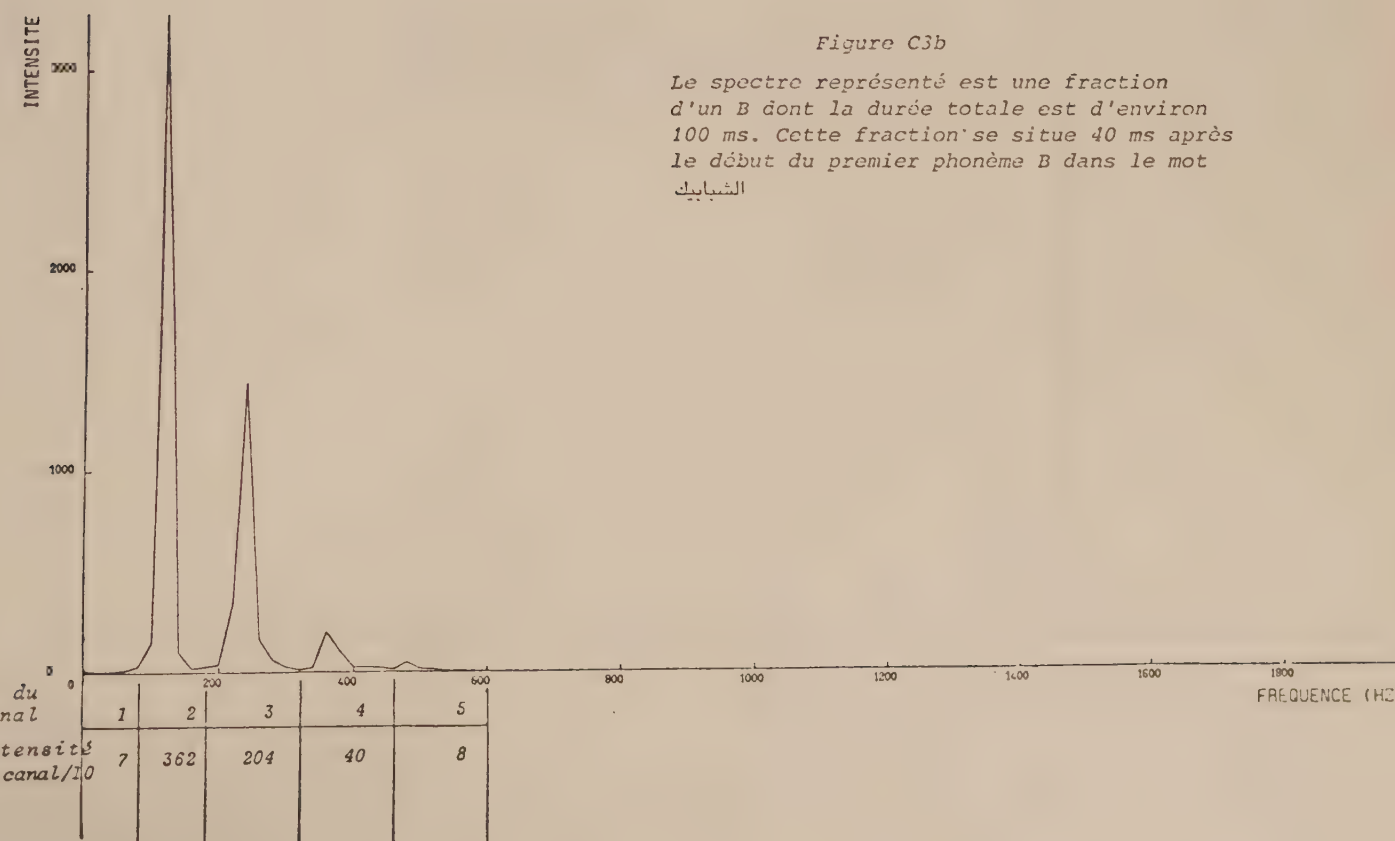


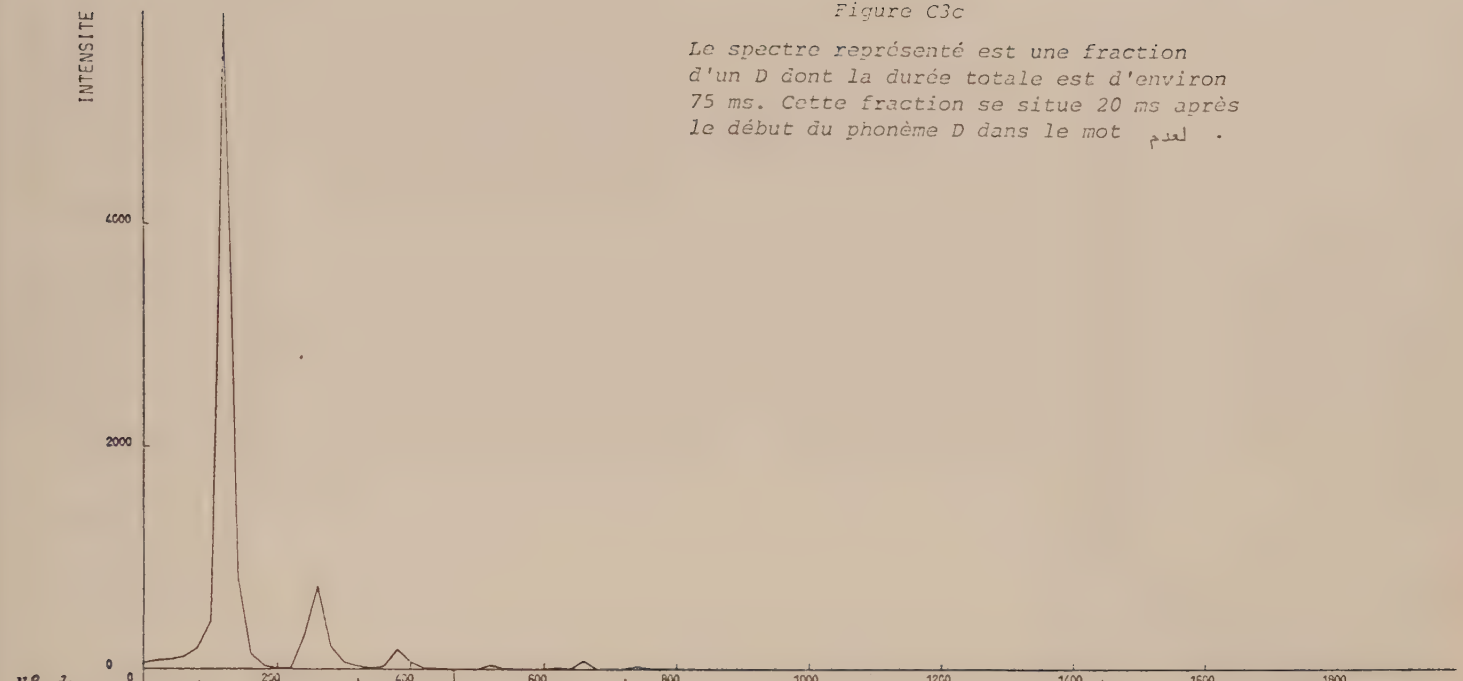
Figure C3b

Le spectre représenté est une fraction d'un B dont la durée totale est d'environ 100 ms. Cette fraction se situe 40 ms après le début du premier phonème B dans le mot الشبابيك

N° du canal	1	2	3	4	5
Intensité du canal/10	7	362	204	40	8

Figure C3c

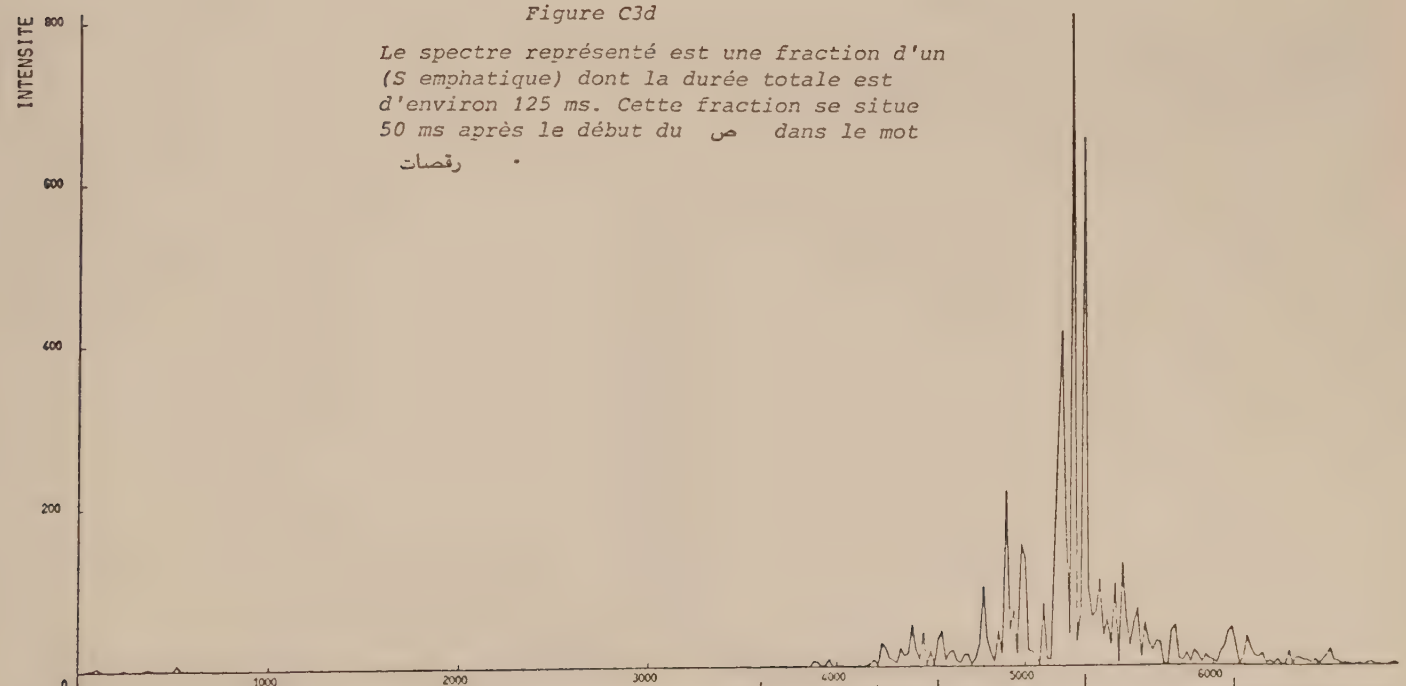
Le spectre représenté est une fraction d'un D dont la durée totale est d'environ 75 ms. Cette fraction se situe 20 ms après le début du phonème D dans le mot لعدم .



N° du canal	1	2	3	4	5	6	7	8	9	10
Intensité du canal/10	53	731	134	29	5	10	4	0	1	1

Figure C3d

Le spectre représenté est une fraction d'un (S emphatique) dont la durée totale est d'environ 125 ms. Cette fraction se situe 50 ms après le début du ص *dans le mot* رقصات .



N° du canal	17	18	19	20	21
Intensité du canal/10	4	2	31	315	207

12	6	9

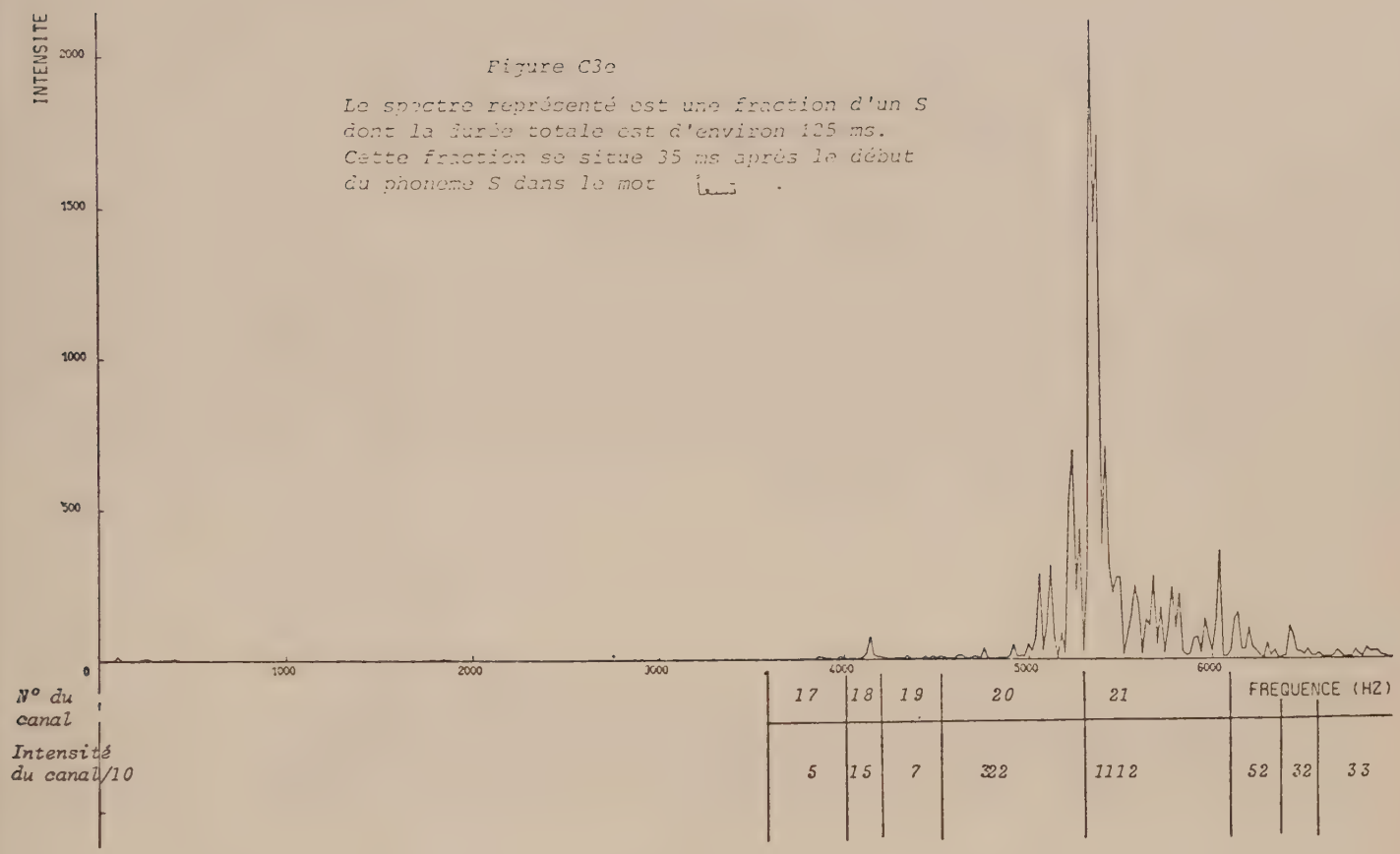


Figure C3e

Le spectre représenté est une fraction d'un S dont la durée totale est d'environ 125 ms. Cette fraction se situe 35 ms après le début du phoneme S dans le mot تسعا .

N° du canal	17	18	19	20	21	FREQUENCE (HZ)		
Intensité du canal/10	5	15	7	322	1112	52	32	33

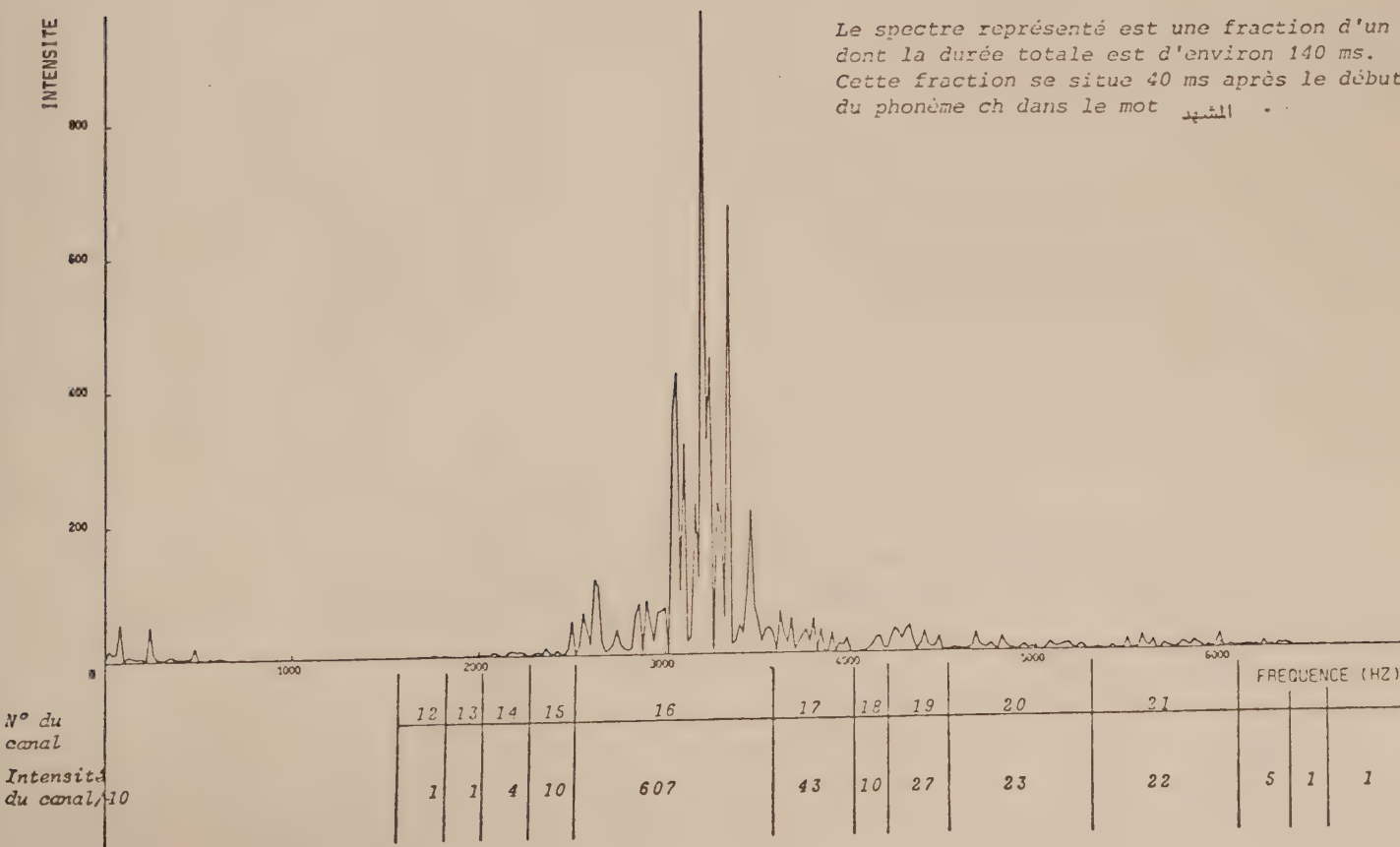


Figure C3f

Le spectre représenté est une fraction d'un ch dont la durée totale est d'environ 140 ms. Cette fraction se situe 40 ms après le début du phonème ch dans le mot المشهد .

N° du canal	12	13	14	15	16	17	18	19	20	21	FREQUENCE (HZ)		
Intensité du canal/10	1	1	4	10	607	43	10	27	23	22	5	1	1

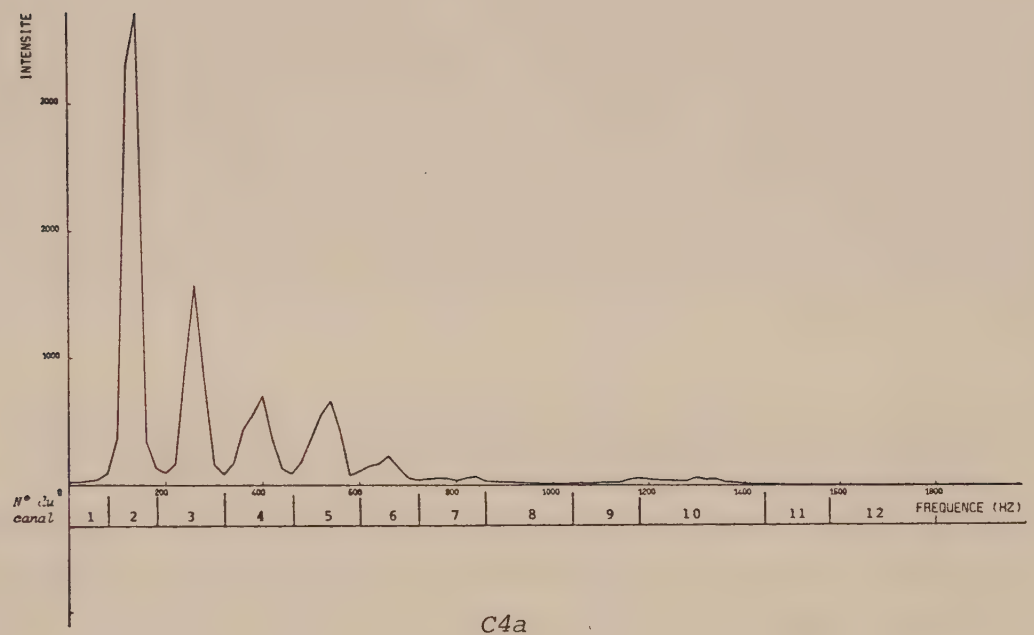
Figures C-4

Intensité moyenne de 500 spectres consécutifs

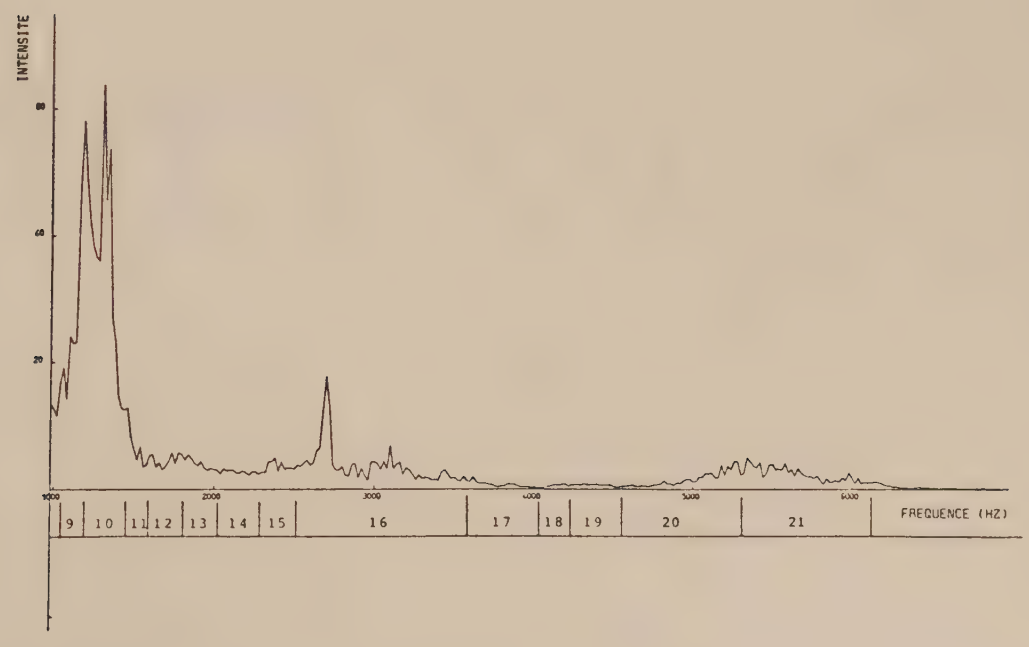
C4a : entre 0 et 2000 Hz

C4b : entre 1000 et 7000 Hz

Les échelles verticales ne sont pas identiques.



C4a



C4b

C.3.2. Découpage en canaux des spectres et calcul des poids

Pour suivre le profil de l'enveloppe spectrale, nous avons découpé le spectre en canaux de nombre et de largeur variables. Le découpage a été effectué d'une part en conciliant empiriquement les spectres des divers phonèmes (figures C3a, b,c,d,e,f) et d'autre part, en faisant le calcul d'un "spectre moyen" par la moyenne arithmétique terme à terme des 500 spectres consécutifs étudiés (figure C4). Les deux méthodes ont abouti à un découpage similaire compte tenu de la limitation à 20 ou 30 canaux qui semblait raisonnable pour assurer une représentation détaillée sans toutefois faire interférer trop les variations de la mélodie (fréquence fondamentale). Un découpage pour des blocs de 1024 points en 25 canaux est décrit dans le tableau C5 .

Le poids d'un canal est évalué en sommant toutes les composantes comprises entre les bornes (inclues). Cependant, pour éviter le caractère parfois arbitraire des bornes et suivre avec plus de finesse le modelé du spectre et pour que la bande passante d'un canal ne soit pas un créneau nettement découpé (figure C6) mais une fonction présentant un large maximum plat, entre une ascension et une descente rapides (figure C7), nous étendons nos canaux en faisant empiéter chacun sur les deux canaux adjacents.

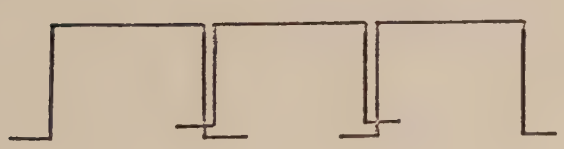


Figure C6

Trois canaux consécutifs à front raide

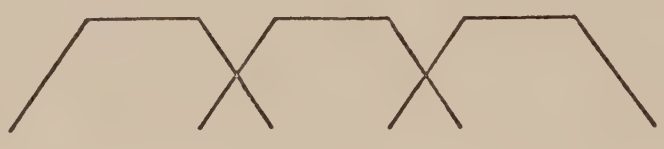


Figure C7

Trois canaux consécutifs à front incliné

Numéro de canal	rang dans le spectre	fréquence borne sup(Hz)	largeur de la bande
1	1-5	80	80
2	6-10	180	100
3	11-17	320	140
4	18-24	460	140
5	25-31	600	140
6	32-37	720	120
7	38-44	860	140
8	45-53	1040	180
9	54-60	1180	140
10	61-73	1440	260
11	74-80	1580	140
12	81-91	1800	220
13	92-102	2020	220
14	103-115	2280	260
15	116-126	2500	220
16	127-180	3580	1080
17	181-202	4020	440
18	203-212	4220	200
19	213-228	4540	320
20	229-266	5300	760
21	267-307	6120	820
22	308-320	6380	260
23	321-329	6560	180
24	330-354	7080	520
25	355-512	10240	3160

Tableau C.5

Description du découpage en 25 canaux

Nous avons défini pour chaque canal c de l'ensemble C, des canaux, une fonction de bande $\phi_c(n)$, approchant la fonction caractéristique, en sorte que

$$\forall \quad n \in [\, 1, 2, \ldots, \frac{N}{2} \,] \qquad \Sigma \{ \phi_c(n) \mid c \in C \} = 1$$

La puissance passant dans un canal c étant :

$$\Sigma \{ | G(\omega_n) |^2 . \phi_c(n) \mid n = n_1 , n_2 \}$$

où n_1 et n_2 sont les nouvelles bornes du canal c.

Une commande PD N1 N2 détermine la fonction ϕ_c et par suite l'élargissement du canal ; N1 et N2 sont deux pourcentages de l'étendue initiale L du canal.

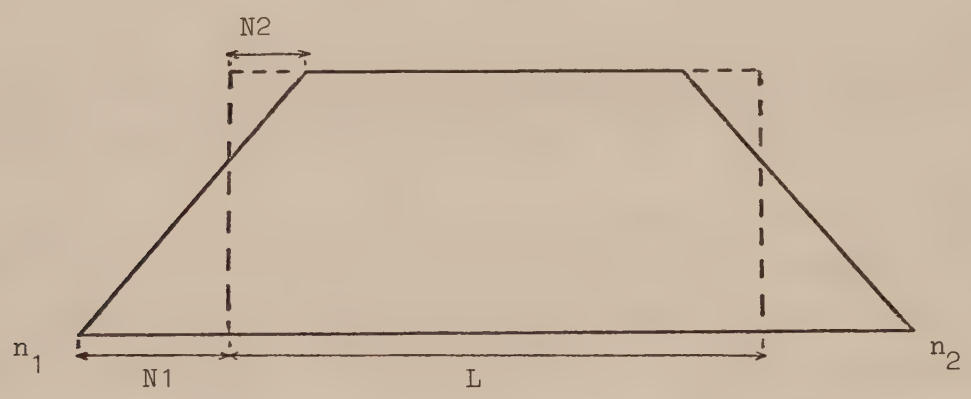


Figure C8

Il est évident que le premier canal n'est pas élargi à gauche, non plus le dernier à droite et que les élargissements ne sont pas forcément symétriques : les étendues initiales des canaux n'étant pas toutes égales.

Nous discuterons plus loin (chapitre E) de l'importance de ces empiètements des canaux et de la stabilité des résultats obtenus avec des paramètres N1, N2 variables.

Nous avons aussi adapté une autre procédure dans le calcul des poids des canaux qui est la suivante :

A partir de la valeur de la mélodie et des composantes du spectre, nous déduisons la courbe enveloppe spectrale et au lieu de faire la somme des composantes du spectre dans une bande déterminée (un canal), nous calculons plutôt l'aire comprise entre cette enveloppe, l'axe des fréquences et les verticales aux bornes du canal car nous pensons que la variation de la mélodie et par suite celles des harmoniques successifs tout au long du discours auront une importance nettement atténuée ce qui rendrait l'analyse moins sensible au choix des bornes des canaux et par conséquent au locuteur.

C.3.3. Les individus (spectres)

Chacun des spectres décalés de 256 en 256 points constitue un individu d'une analyse factorielle. Il convient donc qu'il soit muni d'une étiquette désignant aussi précisement que possible le(s) phonème(s) contenu(s) dans les N points considérés.

Si l'origine de certains phonèmes comme les plosives est nette, dans de nombreux phonèmes, elle est plutôt floue.

Nous avons utilisé des étiquettes comportant 4 symboles, (disponibles sur le clavier de la console et parmi les caractères de l'imprimante et du traceur dont nous disposons), chacun de ces symboles représente un quart du bloc (i.e. pour des blocs de 1024 points, le premier symbole désigne les 256 premiers points, le deuxième les 256 suivants, etc ...). Ainsi l'étiquette AAAB correspond à l'apparition d'un "B" à la suite d'un A dans le dernier quart du bloc. Nous avons utilisé le symbole / pour représenter soit un passage ou une transition entre deux phonèmes, ou sons, soit l'affaiblissement d'un son comme à la fin d'un mot. Nous représentons par des points successifs un son qui "s'éteind" et par un blanc l'absence du signal (surtout au début du signal après un relatif arrêt de phonation).

Pour procéder à l'étiquetage, nous avons observé le signal échantillonné lui-même, mais nous avons surtout identifié les individus par l'audition du son. Pour cela, nous disposons d'une commande PL dans le programme ASPECT permettant la

.../...

restitution du son par conversion numérique - analogique d'un nombre donné de blocs consécutifs commençant ou se terminant, au choix, par un bloc déterminé. Le son est envoyé directement à un casque qui isole l'auditeur des bruits environnants. Quoique longue et fastidieuse, la méthode permet l'étiquetage des blocs avec une précision satisfaisante. Il faut avouer que dans certains cas, les résultats des analyses nous ont permis de confirmer quelques étiquetages douteux ou même de corriger certains autres, toutefois, peu nombreux.

Nous avons choisi de supprimer les silences qui ne pouvaient qu'alourdir le temps de l'analyse sans réellement apporter de contribution appréciable. Nous évitons ainsi d'encombrer la mémoire avec des signaux de bruits indésirables. Pour décider si un spectre correspond à un silence ou non, nous avons fixé à 7 (valeur variable), la limite du poids total du spectre, c'est-à-dire la somme de toutes les composantes contenues dans le spectre. Le poids des spectres varie de 0 à 10000 environ, de sorte que le poids délimitant le silence correspond à 1 0/00 approximativement du son le plus intense. En dépit de ce seuil très bas, le nombre de blocs ainsi éliminés est substantiel puisqu'il est de l'ordre de 16%.

C.3.4. Temps de traitement

L'échantillonnage du signal étant effectué, les différentes étapes du traitement se décomposent de la façon suivante :

- chargement du bloc
- application de la fenêtre de Hanning
- calcul de la transformée de Fourier
- calcul du spectre de puissance
- calcul du poids des C canaux

L'ensemble de ces opérations réclame (toutes opérations d'entrée-sortie inclues) 7,96 secondes pour un bloc de 1024 points, soit avec les conditions optimales que nous avons décrites précédemment, à savoir des blocs de 1024 points décalés de 256 en 256, 100 blocs correspondent à 50 secondes. Ce traitement nécessite donc un temps 160 fois plus long que le temps réel. Néanmoins, les calculs pourraient être faits avec des circuits électroniques appropriés en un temps avoisinant le temps réel avec la technologie actuelle.

CHAPITRE D

ANALYSE FACTORIELLE DES CORRESPONDANCES

Le développement des ordinateurs permet de confronter et de mettre en ordre des ensembles immenses de données. Expérimenter, ce peut être instaurer un système nouveau, autre que celui qu'on se proposait pour objet. L'examen exhaustif de toutes les éventualités, condition nécessaire à la démonstration expérimentale d'une hypothèse, n'est d'ailleurs plus possible quand l'esprit a peine à embrasser l'ensemble des informations.

L'idée de l'analyse des données est de tirer, à partir d'un tableau de données, des résultats qui ne sont autres que ces données, observer des structures, découvrir des phénomènes, réaliser des groupements de faits qui permettent de ne pas regarder toujours les petits détails ; c'est rendre visible ce qui n'est pas globalement visible. C'est un traitement que l'on appelle communément analyse mais qui est bien plutôt synthèse, car il est confrontation de faits nombreux mis ensemble,non résolution d'un tout ordonné, en ses éléments.

Parmi les méthodes d'analyse des données les plus récentes, nous avons choisi l'analyse factorielle des correspondances. De cet "outil pour dégager de la gangue des données le pur diamant de la véridique nature" (Benzécri,1973), nous donnons les idées directrices et une formulation mathématique rapide (inspirées des oeuvres de Benzécri J.P.).

.../...

D.1. CORRESPONDANCE ET TABLEAU DE DONNEES

Deux ensembles, I et J, sont en correspondance s'il existe des relations entre les éléments de ces ensembles et plus précisément si les éléments s'associent par couples (i,j) avec $i \in I$ et $j \in J$. Cette association est exprimée par un nombre réel positif K(i,j). L'ensemble des K(i,j) forme le tableau de données K(I,J) à I lignes et J colonnes.

On nomme généralement individus, les éléments de I et variables ceux de J (bien que les rôles dévolus aux deux ensembles I et J soient rigoureusement symétriques).

On note :

$K(i) = \Sigma \{ K(i,j) \mid j \in J \}$ le total de la ligne i.

$K(j) = \Sigma \{ K(i,j) \mid i \in I \}$ le total de la ligne j.

$K = \Sigma \{ K(i,j) \mid i \in I , j \in J \}$ le total général.

Pour mesurer l'importance relative d'un élément i ou j, on définit :

$$f_i = \frac{K(i)}{K} \text{ (masse de i)} \quad , \quad f_j = \frac{K(j)}{K} \text{ (masse de j)} \quad , \quad f_{ij} = \frac{K(i,j)}{K}$$

d'où les distributions marginales :

$$f_I = \{ f_i \mid i \in I \} \quad \text{pour I}$$

$$\text{et} \quad f_J = \{ f_j \mid j \in J \} \quad \text{pour J}$$

Et, pour se ramener à une correspondance aléatoire (distribution probabiliste), un individu i est caractérisé par son profil sur J

$$f_J^i = \{ f_j^i \mid j \in J \} \text{ où } f_j^i = \frac{K(i,j)}{K(i)} \quad \text{(distribution de J sachant i)}$$

Le profil de j sur I étant :

$$f_I^j = \{ f_i^j \mid i \in I \} \text{ où } f_i^j = \frac{K(i,j)}{K(j)} \quad \text{(distribution de I sachant j)}$$

.../...

D.2. NUAGES ET ESPACES

Un nuage sur J est ainsi un ensemble de card J nombres positifs ou nuls et de total 1 ; donc l'espace des profils sur J est à (card J - 1) dimensions. Dans cet espace, chaque individu i est représenté par son profil f_J^i auquel on affecte la masse f_i. On appelle nuage N(I) et on note :

$$N(I) = \{ (f_j^i , f_i) \mid i \in I \} ,$$

l'ensemble des couples formés du profil d'un individu et de sa masse.

Le centre de gravité de N(I) n'est autre que $f_J = \{ f_j \mid j \in J \}$. De même, le nuage $N(J) = \{ (f_I^j , f_j) \mid j \in J \}$, est défini dans l'espace des profils sur I ; son centre de gravité est

$$f_I = \{ f_i \mid i \in I \}$$

D.3. PROXIMITE ; EQUIVALENCE ET DISTANCES DISTRIBUTIONNELLES

La correspondance entre deux ensembles définit sur chacun d'eux une notion de proximité : si deux individus i et i' s'associent de la même façon aux éléments de J, ils sont proches l'un de l'autre. De même, un élément $i \in I$ est proche des éléments $j \in J$ auxquels il s'associe le plus et vice-versa.

Cette notion de proximité est formalisée par les axiomes suivants :

- La distance entre deux éléments de I est d'autant plus petite que leurs profils sur J sont semblables. Si ces profils sont identiques, la distance est nulle.

- Si deux éléments j et j' de J ont le même profil sur I (ils sont dits équivalents quant à la distribution), on ne change rien aux distances entre les éléments de I en substituant à j et j', un élément unique j_o affecté de la masse

  $$f_{j_o} = f_j + f_{j'}$$ (Principe de l'équivalence distributionnelle).

- La distance ne dépend que des distributions conditionnelles (profils) et cumulées.

.../...

D'où le choix de la distance distributionnelle dite du χ^2, définie sur l'espace des profils sur J où se trouve N(I) :

$$d^2 (f_J^i , f_J^{i'}) = \Sigma \{ \frac{1}{f_j} (f_j^i - f_j^{i'})^2 \mid j \in J \} ,$$

et sur l'espace des profils sur I où se trouve N(J) :

$$d^2 (f_I^j , f_I^{j'}) = \Sigma \{ \frac{1}{f_i} (f_i^j - f_i^{j'})^2 \mid i \in I \}$$

D.4. INERTIE ET VARIANCE

L'inertie d'un nuage $N = \{ (M^i, m_i) \mid i = 1, \ldots, n \}$ par rapport à un point P est la somme des inerties des points (M^i, m_i) par rapport à P.

$$I_P(N) = \Sigma \{ I_p (M^i, m_i) \mid i = 1 , \ldots, n \} = \Sigma \{ m_i \, d^2 (P, M^i) \mid i = 1, \ldots n \}$$

$d^2(P,M^i)$ étant le carré de la distance de P à M^i.

$I_p(N)$ qui caractérise l'écart de N par rapport à P est minimale quand P est au centre de gravité G du nuage.

La variance totale du nuage N est :

$var(N) = I_G(N) / m$ où $m = \Sigma \{ m_i \mid i = 1, \ldots, n \}$ est la masse totale.

La variance totale mesure la dispersion de N autour de son centre de gravité.

Soit Δ une droite passant par G ; l'inertie du nuage le long de Δ ou dans la direction de Δ, est l'inertie par rapport à G de la projection du nuage sur Δ. De même l'inertie du nuage le long ou dans la direction d'un sous-espace passant par G est égale à l'inertie par rapport à G de la projection de N sur ce sous-espace. Cette inertie se décompose en autant de termes qu'il faut d'axes pour engendrer ce sous-espace.

.../...

D.5. AXES PRINCIPAUX D'INERTIE

L'analyse factorielle cherche à découvrir des différences de profils, des intéractions attractives ou répulsives entre les individus et les variables et leur dispersion autour de leurs centres de gravité respectifs f_J et f_I.

Mais pour acquérir une vue sensible de la réalité multidimensionnelle, il importe de réduire chaque nuage à une représentation à la fois accessible à notre vision, donc de dimension aussi faible que possible, et fidèle à la complexité du réel : le nuage s'écarte peu du sous espace sur lequel on projette.

Or, l'écart du nuage, N(I) par exemple, par rapport à un sous-espace L, sera mesuré par l'inertie transversale à L (complémentaire dans I_{f_J} de l'inertie longitudinale dans la direction de L, $I_{//L}(N(I))$. La représentation sera d'autant meilleure que cette inertie sera plus faible ou que sa complémentaire sera maxima.

Le problème est précis : la dimension r étant fixée, $(r \leqslant \text{Inf}\{(\text{card } I-1),(\text{card } J-1)\})$, chercher le sous-espace de dimension r passant par le centre de gravité f_J et rendant maximum l'inertie de N(I) le long de L, $I_{//L}(N(I))$; puis projeter le nuage sur L pour obtenir la représentation approchée désirée.

Le premier axe principal d'inertie est la droite Δ_1 passant par le centre de gravité f_J de N(I) pour laquelle $I_{//\Delta_1}(N(I))$ est maximum. $I_{//\Delta_1}(N(I))$ notée λ_1 est appelée premier moment principal d'inertie. Δ_1 est le premier axe principal d'allongement du nuage, c'est la droite sur laquelle il se projette avec la plus grande dispersion.

Soit E_1, le sous-espace supplémentaire passant par f_J et orthogonal à Δ_1. $N(I)_{E_1}$, étant la projection de N(I) sur E, le deuxième axe principal d'inertie est la droite Δ_2 de E_1, passant par f_J qui maximise $I_{//\Delta_2}(N(I)_{E_1})$ que l'on note λ_2 et que l'on appelle deuxième moment principal d'inertie.

En réalité, pour déterminer les axes principaux d'inertie, on peut se ramener à résoudre un problème algébrique : la diagonalisation d'une matrice carrée symétrique, la recherche des vecteurs propres et valeurs propres de cette matrice ; ce qui fournit à la fois les vecteurs unitaires portés par les axes principaux d'inertie (vecteurs propres) et les moments principaux d'inertie (valeurs propres).

L est le sous-espace engendré par les r premiers axes principaux d'inertie.

D.6. PROPRIETES DES FACTEURS

L'analyse factorielle est un changement d'axes et de coordonnées : le système des Δ_α, axes principaux d'inertie deux à deux orthogonaux est utilisé comme système d'axes de coordonnées, coordonnées qui seront appelées les facteurs.

Au profil f_J^i d'un élément i de I correspondent les coordonnées associées aux r axes principaux d'inertie :

$$f_J^i = \{ F_1(i) , F_2(i) , \ldots F_\alpha(i) , \ldots , F_r(i) \}$$

Le centre de gravité f_J du nuage et (I) étant à l'origine du système, on a :

$$\forall \alpha \quad \Sigma \{ f_i F_\alpha(i) \mid i \in I \} = 0 :$$

- *Les fonctions F_α ont une moyenne nulle*

L'inertie du nuage le long de l'axe α est :

$$I_{//\Delta_\alpha}(N(I)) = \lambda_\alpha = \Sigma \{ f_i F_\alpha^2(i) \mid i \in I \}$$

- *λ_α apparaît ainsi comme la variance de F_α*

- *Les facteurs sont orthogonaux :*

$$\forall \quad \alpha \neq \beta \rightarrow \Sigma \{ f_i F_\alpha(i) F_\beta(i) \mid i \in I \} = 0$$

Donc deux facteurs distincts ont une covariance nulle et ne sont pas corrélés.

.../...

D.7. CONTRIBUTIONS

L'inertie totale du nuage N(I) par rapport à son centre de gravité f_J est :

$$I_{f_J}\ (\ N(I)\ ) \quad = \quad \Sigma\ \{\ f_i\ d^2\ (f_J^i\ ,\ f_J)\ \mid\ i \in I\ \}$$

$$= \quad \Sigma\ \{(f_{ij} - f_i\ f_j)^2\ /\ (f_i f_j)\ \mid\ i \in I\ ;\ j \in J\}$$

mais aussi :

$$I_{f_J}\ (\ N(I)\ ) \quad = \quad \Sigma\ \{\ \lambda_\alpha\ \mid\ \alpha = 1,\ \ldots,\ r\ \} = T = \text{Trace}$$

où r est le nombre total des axes factoriels du nuage.

La contribution relative de l'axe α à l'inertie totale du nuage, ou son taux d'inertie est :

$$\tau_\alpha \quad = \quad \lambda_\alpha\ /\ T$$

La contribution relative de l'élément à cette inertie totale est :

$$\text{INR}\ (i) = f_i\ d^2\ (f_J^i\ ,f_J)\ /\ \Sigma\ \{\ f_i\ d^2\ (f_J^i\ \ ,\ f_J)\ \mid\ i \in I\ \}$$

λ_α inertie du nuage le long de l'axe α est l'inertie par rapport à f_J de la projection du nuage sur cet axe donc la somme des inerties des points de cette projection. Ainsi, la contribution absolue de l'élément i à λ_α ou à l'axe α est :

$$f_i(F_\alpha(i))^2$$

Alors la contribution relative de l'élément i à λ_α est :

$$\text{CTR}_\alpha(i) \quad = \quad f_i(F_\alpha\ (i)\ )^2\ \ /\ \ \Sigma\ \{\ f_i\ (F_\alpha\ (i)\ )^2\ \mid\ i \in I\ \}$$

La valeur de cette contribution permet de reconnaître si un axe α résulte de quelques points seulement ou d'un apport équilibré des divers points du nuage.

.../...

— L'inertie d'un point f_J^i par rapport au centre f_J est :

$$I_{f_J}(f_J^i) = f_i \, d^2(f_J^i, f_J)$$

$$= f_i \, \Sigma \{ (F_\alpha(i))^2 \mid \alpha = 1, \ldots, r \}$$

$$= \Sigma \{ f_i \, (F_\alpha(i))^2 \mid \alpha = 1, \ldots, r \}$$

ce qui fournit une décomposition de l'inertie d'un point f_J^i. La part relative à l'axe α que l'on note $COR_\alpha(i)$ est :

$$COR_\alpha(i) = f_i \, (F_\alpha(i))^2 \; / \; \Sigma \{ f_i (F_\alpha(i))^2 \mid \alpha = 1, \ldots, r \}$$

$$= (F_\alpha(i))^2 \; / \; \Sigma \{ (F_\alpha(i))^2 \mid \alpha = 1, \ldots, r \}$$

$$= (F_\alpha(i))^2 \; / \; d^2(f_J^i, f_J)$$

$COR_\alpha(i)$ s'interprète comme le carré du cosinus de l'angle que forme le rayon (f_J, f_J^i) avec l'axe Δ_α (aussi comme le carré d'un coefficient de corrélation). La considération de $COR_\alpha(i)$, pour les divers axes, permet de savoir de quels axes f_J^i est le plus voisin ou quels facteurs expliquent la position de f_J^i relativement au centre du nuage.

— La qualité de la représentation d'un point sur un sous-espace engendré par les axes α, β, ... γ par sa projection, est exprimée par le rapport du carré de la distance à f_J de la projection de f_J^i sur ce sous-espace au carré de la distance de f_J^i à f_J :

$$d^2(pr_{(\alpha,\beta,\ldots,\gamma)}(f_J^i), f_J)^2 \; / \; d^2(f_J^i, f_J) = COR_\alpha(i) + COR_\beta(i) + \ldots + COR_\gamma(i)$$

Plus cette valeur est voisine de 1, meilleure est la représentation. D'ailleurs, ce rapport n'est autre que le carré du cosinus de l'angle en f_J du triangle $(f_J, f_J^i, pr_{(\alpha,\beta,\ldots,\gamma)}(f_J^i))$.

On note généralement QLT, la qualité de la représentation correspondant au nombre de facteurs calculés.

.../...

D.8. FACTEURS SUR LES DEUX ENSEMBLES - FORMULES DE TRANSITION ET DE RECONSTITUTION

Les ensembles I et J mis en correspondance par le tableau K(I,J) jouent des rôles parfaitement symétriques. On peut prendre l'ensemble J décrit par l'ensemble I et obtenir ainsi le nuage N(J) dans l'espace des profils sur I, puis faire l'analyse de ce nuage.

Or, dans toute analyse factorielle de correspondances, la même suite de valeurs propres est issue de l'un et de l'autre des deux nuages N(I) et N(J) ; d'où l'identification des axes de même rang, donc des sous-espaces de représentation engendrés par ces axes, et la superposition des deux nuages et de leurs supports. C'est dans ce nouveau sous-espace de dimension r ($r \leq$ Inf {(card I-1) , (card J-1)}) que seront représentés à la fois les deux ensembles I et J.

En notant G_α, les facteurs extraits de N(J), on a les deux formules dites de transition (des facteurs sur I ou facteurs sur J et réciproquement) qui manifestent la parfaite symétrie des rôles dévolus en analyse des correspondances aux deux ensembles I et J :

$$F_\alpha(i) = (1/\sqrt{\lambda_\alpha}) \ \Sigma \{ f_J^i \ G_\alpha(j) \mid j \in J \}$$

$$G_\alpha(j) = (1/\sqrt{\lambda_\alpha}) \ \Sigma \{ f_i^j \ F_\alpha(i) \mid i \in I \}$$

$F_\alpha(i)$ apparaît donc, au coefficient $1/\sqrt{\lambda_\alpha}$ près, comme la moyenne des $G_\alpha(j)$ pondérés par les masses f_J^i pour j parcourant J.

De même $G_\alpha(j)$ apparaît, au coefficient $1/\sqrt{\lambda_\alpha}$ près, comme la moyenne des $F_\alpha(i)$ pondérés par les masses f_i^j pour i parcourant I.

On dit ainsi que les i sont les barycentres (corrigés par, ou homothétiques dans, le rapport $1/\sqrt{\lambda_\alpha}$) des j et les j sont les barycentres (corrigés par, ou homothétiques dans, le rapport $1/\sqrt{\lambda_\alpha}$) des i.

.../...

Le principe barycentrique permet, dans la représentation simultanée de N(I) et N(J) de découvrir des intéractions attractives ou répulsives entre les individus et les variables, de rendre raison des groupements de points i et j mêlés, qui se voient à la périphérie du nuage : dans un tel groupement mixte, les éléments j fournissent la raison commune de la présence des éléments i et réciproquement la présence des j s'explique par l'attraction des i. En particulier, les groupements qui s'opposent aux extrêmités d'un axe facilitent son interprétation et permettent souvent de lui associer une qualification générale.

Les formules de transition présentent l'intérêt de pouvoir calculer les facteurs sur l'un des deux ensembles en utilisant l'analyse de l'autre.

La connaissance des valeurs des facteurs sur I et J, complétée par celle des valeurs propres λ_α et des profils moyens f_I et f_J, permet de reconstituer le tableau initial des données :

$$K(i,j)/K = f_{ij} = f_i\, f_j \left(1 + \Sigma \{ \lambda_\alpha^{-1/2} F_\alpha(i)\, G_\alpha(j) \mid \alpha = 1, \ldots, r \} \right)$$

C'est la formule dite de reconstitution des données en fonction des facteurs (formule qui est *exacte* si on tient compte de tous les facteurs , mais n'est qu'*approchée* si on se borne aux premiers facteurs).

D.9. ELEMENTS SUPPLEMENTAIRES

On appelle élément supplémentaire, tout élément i ou j adjoint à l'analyse tenant compte exclusivement des autres éléments de I et de J, dits éléments principaux.

En effet, tout profil π_J peut être projeté sur l'espace de représentation de N(I) donc sur les axes factoriels issus de l'analyse du nuage N(I). De même, tout profil π_I peut être projeté sur l'espace de représentation de N(J) donc sur les axes factoriels issus de l'analyse du nuage N(J).

.../...

Ceci présente l'intérêt parmi d'autres, de pouvoir situer un individu (ou une variable) nouveau par rapport aux autres déjà étudiés en calculant pour lui les valeurs des facteurs sans pour autant refaire l'analyse ou lui donner de part à la constitution des axes.

Enfin, c'est la même formule de transition du cas général qui s'applique au cas d'un élément supplémentaire.

CHAPITRE E

ANALYSE DE LA PAROLE

E.1. REDONDANCE

Le propre des messages sensoriels est qu'ils transmettent d'autant plus d'informations et d'autant mieux qu'ils sont plus redondants. Le message vocal est très redondant, qualité qui est assurée notamment par la stabilité de certaines de ses zones.

Il est donc intéressant de mettre à profit cette redondance pour l'analyse de la parole, de façon à diminuer les volumes des données à traiter et, par suite, le temps de traitement.

Ainsi, sommes-nous conduits à procéder à une paramétrisation, une compression et un codage du signal.

E.2. PARAMETRISATION ET CODAGE

On distingue deux types de traitement du signal de la parole : les méthodes temporelles et les méthodes spectrales ou fréquencielles.

E.2.1. Méthodes temporelles

Ce sont des méthodes qui travaillent directement sur le signal. Parmi les méthodes temporelles, on peut citer :

- les méthodes de passage par zéro. De façon générale, elles consistent à filtrer le signal dans plusieurs bandes de fréquence pour tenter de séparer les formants ; puis on compte les passages par zéro dans chaque bande et leurs durées.

- les méthodes du codage des extrêma du signal.

- les analyses en fonction d'ondes : l'examen du signal de parole filtré dans une bande assez étroite ressemble à une sinusoïde modulée par une enveloppe.

.../...

- les méthodes de prédiction linéaire : elles reposent sur l'hypothèse que tout échantillon du signal de parole peut être prédit sous la forme d'une combinaison linéaire d'un certain nombre d'échantillons précédents (Markel J.D. and Gray A.H. 1976).

Nous ne discutons pas les avantages et les inconvénients de ces méthodes, le choix n'étant toujours pas tranché. Toutefois, on trouve un examen comparé de quelques méthodes temporelles dans Maïssis A. (1973).

E.2.2. Méthodes spectrales

Nous choisissons la représentation spectrale pour des raisons précises (cf. § C.2.2.1.) et d'autant plus que les données peuvent être soumises à l'analyse factorielle qui nous paraît un excellent outil de compression et de codage.

En effet, la représentation en fréquence du signal de la parole est avantageuse à, au moins, deux points de vue : d'une part, l'étude de l'audition montre que l'oreille effectue dans un premier temps une sorte d'analyse en fréquence, d'autre part, cette forme de représentation spectrale du signal vocal dans le domaine des fréquences est simultanément concise et fidèle.

Comme la paramétrisation de la parole consiste à extraire du signal un nombre restreint de paramètres pertinents, tout en éliminant en partie l'information redondante et en gardant le maximum d'informations utiles (il est de toute façon impossible de séparer exactement information utile et information redondante dans la parole), nous commençons par déterminer la puissance du spectre dans un certain nombre de bandes de fréquence dites canaux.

L'idée de ne pas transmettre le spectre lui-même mais le spectre d'énergie n'est pas nouvelle, comme en témoigne le vocodeur (Voice Coder) à canaux qui a déjà une très longue histoire ; dès 1930, Dudley W. en avait défini les principes.

Le succès relatif de cette méthode, même appliquée avec une définition spectrale grossière, est la principale justification de l'expérience que nous avons tentée : se prévaloir de toute la précision et de la souplesse des méthodes digitales pour l'analyse spectrale complète de la parole.

.../...

E.2.3. Codage spatial

Nous suivons, par notre choix, le principe suivant : pour traiter de la qualité, il faut traduire au préalable, en termes mathématiques, les relations au sein du système dont on veut saisir la structure qualitative. Nous disons que c'est, déjà, un problème de codage.

Nous avons défini (cf. § C.2 et § C.3), l'ensemble I des individus qui résulte d'un découpage aveugle de la chaîne de parole et l'ensemble J des canaux qui n'est qu'une approximation discontinue -on dit parfois une discrétisation- de l'intervalle continu de toutes les fréquences.

Notre codage sera essentiellement un codage spatial. "Puisqu'il faut confier aux mathématiques les relations observées entre des êtres linguistiques, on choisira la géométrie : province des mathématiques où, par l'imagination, tout esprit peut trouver accès. C'est dans l'espace, qu'entre les faits actuellement observés, apparaît le mieux la place des faits possibles : dans l'espace, l'actuel sous-tend le potentiel. Mais comme il est de règle depuis Descartes, nous nous élevons sur l'échafaudage des nombres : comment pénétrer autrement, un espace aux multiples dimensions ?".

On traduira donc les similitudes entre les éléments de l'ensemble I par des proximités entre points d'un espace E de dimension C (nombre de canaux) : à deux éléments i et i' semblables dans I correspondront deux points proches dans E muni d'une distance, vérifiant le principe d'équivalence distributionnelle, propre à l'analyse des correspondances.

E.2.4. Statistique sans probabilité ?

Nous sommes de l'avis de Benzécri J.P. pour une statistique sans probabilité : le modèle probabiliste a une grande valeur suggestive, mais on ne peut pas l'ajuster exactement au réel. En particulier, les épreuves de validité sont douteuses. D'ailleurs, il n'est jamais sage de poser a priori les hypothèses précises d'un modèle auquel on réfère ensuite les faits ; résolu à n'abandonner le modèle que si du point de vue statistique, il apparaît inacceptable. Nous préférons procéder inductivement : demander à l'ordre des faits observés qu'il suggère lui-même des hypothèses voire un modèle, c'est ce que permet l'analyse factorielle d'un tableau de données.

.../...

E.2.5. Réduction des données

Pour découvrir des structures et reconnaître des formes, nous envisageons, après notre premier codage spatial, le problème de la compression ou réduction des données et donc de l'information par le passage de l'espace E à un sous-espace optimal. De plus, il nous a paru essentiel d'utiliser un opérateur de réduction indépendant des paramètres utilisés et distinct des procédures de discrimination qui doivent suivre cette opération.

L'analyse factorielle des correspondances, pour bien d'autres raisons, nous a semblé être la plus intéressante des méthodes de réduction des données : suivre directement une trajectoire dans l'espace des variables est impossible mais en se bornant à caractériser le profil spectral par les premiers facteurs, la tâche devient possible.

De plus, les premiers facteurs décrivent bien le profil : si l'on calcule celui-ci à partir de ceux-là par la formule fondamentale de reconstitution, on doit produire un signal sonore qui pour l'oreille ne diffère guère du signal produit d'après le profil lui-même ; ce que nous projetons de vérifier.

Parmi les équipes qui ont traité le signal de parole par l'analyse factorielle (pas nécessairement celle des correspondances), nous pouvons citer celle de Pols (1969)(1973) et également celles de Li, Hughes et House (1973) et de Graillot (1973).

E.3. ANALYSES DES TABLEAUX K(I,J)

E.3.1. Influence de la durée de l'échantillon de parole

La lecture, dans des conditions courantes de prononciation (c'est-à-dire avec un débit normal et une articulation naturelle), du texte choisi pour servir de premier support à notre travail a duré environ quatre minutes.

En fixant à 20 480 Hz la fréquence d'échantillonnage, ces quatre minutes de parole correspondent à près de cinq millions de nombres, soit cinq mille spectres de 1024 points (de durée 50 ms chacun) ou dix mille spectres de 512 points (de durée 25 ms chacun).

.../...

Nous découpons le signal en tranches successives A_1, A_2, ..., A_{10} de 500 spectres de 1024 points (la durée de la tranche est donc de 25 s) et en tranches, toujours successives, B_1, B_2, ..., B_{20} de 500 spectres de 512 points (la durée de la tranche est 12,5 s).

Pour juger de la variabilité ou de la stabilité des résultats en fonction du texte étudié, nous avons soumis à l'analyse des correspondances, les tableaux $K_m(I,J) \mid m = 1, \ldots,10$ afférents aux dix tranches A_1, ..., A_{10}, mais aussi les tableaux construits d'après deux, trois, quatre ou cinq tranches successives ou non.

Nous avons procédé de même pour les tableaux $K_n(I,J) \mid n = 1, \ldots,20$ des tranches $B_n \mid n = 1, \ldots, 20$.

Notre analyse de référence ne pouvait être que celle du tableau K(I,J) correspondant à tout le texte, résultant de la juxtaposition verticale des dix tableaux $K_m(I,J)$ des $A_m \mid m = 1, \ldots, 10$ ou des vingt tableaux des $B_n \mid n = 1,\ldots,20$.

Des critères de base retenus, nous citons :

- Le nombre de facteurs "significatifs" extraits d'une analyse.
- L'inertie représentée par l'ensemble de ces facteurs.
- La présence des facteurs qui apparaissent comme des phénomènes généraux reflétant une structure permanente des données.
- L'importance de chacun de ces facteurs permanents et donc le classement de ces facteurs pour chacune des analyses.

Nos observations sont les suivantes :

- Un signal de parole d'une durée de 12,5 s n'est pas, en général, porteur de l'information suffisante pour découvrir la structure propre d'un langage, surtout s'il correspond à une tranche d'un texte choisi arbitrairement sans aucune hypothèse ou contrainte, l'analyse factorielle ne faisant que dégager l'essentiel de l'information présente dans le tableau de données qu'elle traite.

- Les analyses des tableaux $K_m(I,J)$ (relatifs à une durée de 25 s) donnent des résultats analogues à ceux du grand tableau K(I,J) correspondant à la durée des

.../...

quatre minutes, avec toutefois pour quelques uns, des permutations entre les axes : par exemple, les facteurs trois et quatre de l'analyse d'un tableau $K_m(I,J)$ sont les facteurs quatre et trois de l'analyse du tableau K(I,J). Ceci n'altère pas, à nos yeux, la stabilité d'une telle analyse parce qu'il correspond à une plus grande présence dans les données d'une tendance par rapport à une autre : l'essentiel étant de retrouver toutes les tendances ou les structures propres au langage, sans en perdre aucune.

Pour la suite de nos travaux, nous avons choisi une tranche de signal d'une durée de 25 s, sans omettre de comparer, quand il le faut, les résultats obtenus à ceux d'une durée plus grande.

Ceci dit, et après l'examen des analyses des tableaux $K_n(I,J) \mid n=1,...,20$, il paraît possible de choisir ou plutôt de construire un texte d'une durée de lecture inférieure à 25 s qui pourrait servir d'échantillon de référence. On mettrait ainsi en place les axes de l'analyse, sinon la place même des zones afférentes à chaque phone, lesquelles ne peuvent être délimitées sans un échantillon d'au moins 25 s. Malgré l'intérêt qu'un tel texte présente, nous laissons pour plus tard ce travail.

E.3.2. Stabilité des résultats en fonction de l'échantillon de parole

Notre texte, rappelons-le, est lu d'une manière naturelle sans aucun effort particulier d'accentuation ou d'intonation, avec un débit normal ; ceci n'implique pas une régularité d'articulation et ne résoud pas le problème de la variabilité qui caractérise la parole.

En effet, bien que le signal de parole puisse être caractérisé par les domaines de définition de ses paramètres, il est d'une grande variabilité à l'intérieur de ces domaines, lesquels ne sont pas tous parfaitement connus d'ailleurs (HATON, 1974).

Une même unité phonétique a des réalisations physiques différentes suivant le contexte dans lequel elle se trouve et suivant le locuteur. Il y a un nombre important d'altérations possibles.

On pourra juger de la qualité d'une paramétrisation ou d'un codage par sa faible variation en fonction de l'évolution dans le temps et par suite, en fonction du texte,de l'environnement et du locuteur.

.../...

Quoique notre texte soit lu d'une manière presque uniforme, notre signal de parole est sûrement entaché d'un certain nombre de variations dont il est intéressant de connaître les conséquences. Or les analyses des tableaux $K_m(I,J)$ pris seuls, ou juxtaposés deux à deux, ..., cinq à cinq, pour des tranches successives ou non dans le temps, conduisent à des résultats stables et confirment que notre codage par l'analyse factorielle est, tout au moins, très peu sensible aux variations inhérentes à notre signal qui sont difficiles à cerner et à déterminer. Ce fait est rassurant quant aux résultats obtenus en ne travaillant que sur une seule tranche de 25 s.

Ceci n'exclut pas qu'il reste à examiner les conséquences d'autres variations possibles (liées au locuteur) sur les résultats acquis.

E.3.3. Suppression du silence

Dans la continuité du fonctionnement mécanique de l'appareil vocal, on distingue deux états qui correspondent à des moments de silence durant le discours ; un silence qui n'est d'ailleurs jamais parfait.

Le premier état est celui qui précède l'occlusion et la libération énergétique des plosives se trouvant au début d'un mot (et donc de durée difficile à déterminer) ou à l'intérieur d'un mot (sa durée est généralement inférieure à la durée de nos spectres sous la réserve qu'elle dépend évidemment du débit de parole...).

L'autre silence correspond aux courtes pauses indispensables qui peuvent, -mais rarement- arriver dans la prononciation d'un mot.

Nous avons décidé de supprimer de notre tableau de données toutes les lignes (i.e. les spectres) qui ont une faible énergie (total de la ligne) en fixant un seuil déterminé après examen de l'ensemble des lignes.

Pour notre tableau de référence sur les 2000 spectres, nous en gardons 1674. Nous éliminons donc environ 16% du temps de signal (cf. § C.3.2.).

De la comparaison des analyses des deux tableaux, nous déduisons que la suppression du silence, avec le bénéfice de la réduction du temps de traitement, n'affecte en aucun cas le codage des plosives dans l'espace des axes factoriels

.../...

retenus et d'autre part permet une meilleure clarté dans le plan des facteurs six et sept où s'expriment particulièrement quelques sons comme le H aspiré dont l'énergie est localisée dans les basses fréquences.

Nous observons aussi que d'autres sons de faible énergie comme le F sont mieux repérés et codés.

E.3.4. Choix de l'empiètement des canaux

Nous écrivions au § C.3.1. que pour diverses considérations relatives au découpage du spectre en canaux, nous étendons les canaux en faisant empiéter chaque canal sur les deux canaux adjacents. Le pourcentage d'élargissement de la bande d'un canal est le même pour tous les canaux, il est défini par la commande PD (cf. § 3.1.) qui calcule la puissance de chacun des canaux élargis sans modifier la puissance totale du spectre et ceci en exigeant des empiètements symétriques par rapport aux bornes délimitant initialement les canaux (Figure E.1).

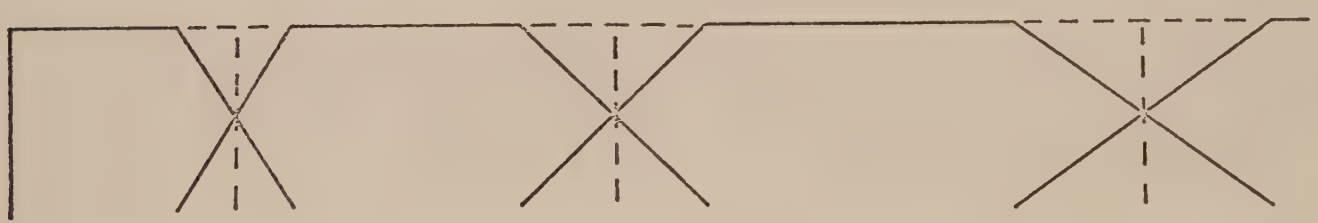


Figure E.1. : Canaux élargis de 25%

.../...

Mais les largeurs des bandes des canaux sont inégales ; de plus, elles sont importantes dans les hautes fréquences et faibles dans les basses. Ceci implique une limitation du taux d'élargissement des canaux.

Nous retenons trois analyses avec les élargissements respectifs 0% (sans empiètement), 20% et 30%.

Les deux premières analyses conduisent aux mêmes résultats : les sept premiers axes factoriels sont, dans leur ensemble, les mêmes et totalisent presque la même inertie (89.089% et 89.35 % respectivement).

Cependant, avec un empiètement de 20%, les canaux C01 à C16 couvrant la bande (0 - 3580 Hz) ont presque tous une meilleure qualité de représentation QLT (voir tableau E2) et tout particulièrement les canaux C11, C12 et C13 (bande 1440, 1580, 1800, 2020 Hz). Ceci se traduit, au niveau des groupes de sons par une meilleure homogénéité de paramétrisation (moindre dispersion des valeurs des facteurs) pour certains et une meilleure représentation (plus grande CTR) pour d'autres comme le $\mathcal{C}$.

La baisse de QLT observée généralement pour les canaux de haute fréquence est dûe à la grande largeur de bande de ces canaux où se localise, presque exclusivement, l'énergie des sons comme S, X (ch) ... qui sont toujours bien repérés et représentés.

Avec un empiètement de 30%, ce sont ces derniers canaux qui subissent les plus grandes variations et malgré cela ce sont les mêmes six premiers axes factoriels qui apparaissent, le septième nouvel axe n'étant pas significatif.

La stabilité des résultats avec un empiètement de 20% ne nous a pas surpris : les travaux de notre Thèse de 3ème cycle (Moussa T., 1972) permettaient de la prévoir.

Nous pensons que le meilleur empiètement doit se situer autour de 20% et qu'il dépend très probablement du locuteur.

.../...

Canal	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	C11	C12
QLT empiètement 0%	996	1000	1000	997	1000	996	998	684	983	989	382	402
QLT empiètement 20%	1000	1000	1000	999	1000	997	999	740	971	999	427	700

C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23	C24	C25
393	738	931	1000	970	788	881	1000	1000	911	766	934	782
734	755	960	1000	962	734	880	1000	998	849	739	918	782

TABLEAU E.2.

E.3.5. Sous-canaux supplémentaires

Nous avons expliqué au § C.3.1. comment découper le spectre en canaux de largeur variable et ceci à partir d'un spectre moyen et des spectres des divers phonèmes.

Or, comme l'indique la figure C.4. du spectre moyen, le choix du nombre et des frontières des canaux, dans la bande 0 à 1600 ou 2000 Hz environ, ne posait aucun problème. Par ailleurs, c'est dans la zone des hautes fréquences, domaine des X(ch), S, ʃ (S emphatique propre à l'arabe) ... que le choix était plus élaboré.

En effet, au départ, nous avions cru bon d'avoir un découpage aussi fin que dans la bande des faibles et moyennes fréquences. Mais l'analyse avec un découpage du spectre en 32 canaux révélait les inconvénients suivants :

- Elaboration de facteurs non significatifs car ils se caractérisent par des oppositions entre des spectres d'un même phonème par exemple le X (ch).
- Très fortes dispersions des coordonnées des spectres dans chacun des groupes de phonèmes à forte énergie dans les hautes fréquences, ce qui poserait, ultérieurement, des problèmes de reconnaissance de spectre ou de phonème.

.../...

Nous expliquons ce résultat par les observations suivantes :
L'essentiel de l'énergie des sons de haute fréquence, contrairement à ceux de faible et moyenne fréquences, se déplace, selon le contexte phonétique du phonème, dans une bande relativement large.

Pour parer à ces inconvénients, nous étions amenés à choisir des canaux assez larges, parfois même très larges, pour couvrir la bande des hautes fréquences (voir tableau C.5.), puis nous avons découpé ces canaux larges en des sous-canaux que nous avons mis en variables supplémentaires (le 16ème en cinq sous-canaux, le 17ème en deux, le 20ème en cinq, le 21ème en quatre) et ceci pour deux raisons :
- avoir une reconstitution plus précise des données à partir des valeurs des facteurs, par la dite formule de reconstitution et ceci dans le cas d'une éventuelle utilisation de l'analyse des correspondances comme moyen de compression dans une transmission de données.
- avoir la possibilité d'étudier les variations des spectres d'un même phonème en fonction de leurs affinités ou de leurs attractions par certains sous-canaux et essayer d'expliquer ces variations par le contexte phonétique.

E.3.6. Analyses : Résultats et Comparaisons

Nous considérons la seconde tranche A_2, de durée 25 ms du signal de parole, dont le profil phonémique et les résultats d'analyse représentent ceux de tout le signal (de durée quatre minutes environ).

Nous commentons et comparons les trois analyses principales suivantes :

- *Analyse 1* de 817 spectres successifs de 25 ms décrits par 25 canaux (déduits du tableau C.5.) où l'intensité d'un canal est évaluée en sommant toutes les composantes comprises entre les bornes du canal.
- *Analyse 2* de 1674 spectres successifs de 50 ms décalés de 12,5 ms (chevauchant de 37,5 ms). Les variables sont de même nombre et de même nature que dans l'analyse 1.
- *Analyse 3* Les spectres sont ceux de l'analyse 2 mais le poids d'un canal est l'aire comprise entre l'enveloppe spectrale, l'axe des fréquences et les verticales aux bornes du canal.

.../...

E.3.6.1. Les facteurs stables et leurs inerties

Les six premiers facteurs sont stables et expriment identiquement la même structure des données décrites par trois tableaux différents. Cette stabilité est visible sur les tableaux E.4. où les facteurs de même rang sont dûs aux mêmes canaux, de plus forte contribution, pour chacune des trois analyses.

Les tableaux E.3. donnent, pour chaque analyse, les inerties simples et cumulées (exprimées en pourcentage) de chaque facteur. Les six premiers totalisent respectivement 85.42, 87.04 et 90.09 % de l'inertie du nuage, ce qui est déjà rassurant quant à la quantité d'information portée par ces facteurs.

Précisons que, pour des considérations acoustiques, nous limitons aux treize premiers canaux (bande de fréquence : 0 - 2020 Hz) l'utilisation de l'enveloppe spectrale approximée dans le calcul des poids. Pour les autres canaux, le poids est la somme des composantes du spectre.

L'utilisation de l'enveloppe spectrale modifie considérablement le poids du premier canal (0 - 80 Hz) qui a une contribution non négligeable au 7ème facteur de l'Analyse 1 ; à ce facteur est substitué un facteur non significatif. Ceci nous a conduit à regrouper par la suite les deux premiers canaux en un seul.

E.3.6.2. Caractérisation du signal par des zones de fréquence

Les tableaux E.4. donnent pour chaque analyse la qualité de représentation, le poids relatif et l'inertie de chaque variable (canal), son abscisse, sa corrélation et sa contribution pour chaque facteur.

L'observation de ces tableaux nous permet, par regroupement des canaux de fortes contributions et de même signe d'abscisse sur un axe, de déterminer pour chaque facteur les zones de fréquence qui le caractérisent.

Le premier facteur partage le spectre en deux zones :

. Z.1.1. De zéro à ∿ 1800 Hz

. Et Z.1.2. de 1800 à ∿ 7500 Hz

TITRE: ANALYSE 1

LES VALEURS PROPRES

NUM	ITER	VAL PROPRE	POURCENT	CUMUL	I*I HISTOGRAMME DES VALEURS PROPRES DE LA MATRICE
2	1	.81362832	32.360	32.360	I*I**********I**********I**********I**********
3	1	.59145319	23.524	55.884	I*I**********I**********I**********
4	1	.33759153	13.427	69.311	I*I**********I**********
5	1	.17358151	6.904	76.214	I*I**********
6	1	.13676763	5.440	81.654	I*I**********
7	1	.09486625	3.773	85.427	I*I*******
8	0	.05334372	2.122	87.549	I*I****
9	1	.04693530	1.867	89.415	I*I***
10	2	.04221806	1.679	91.094	I*I***
11	2	.04011095	1.595	92.690	I*I***
12	2	.03728659	1.483	94.173	I*I***
13	3	.03548246	1.411	95.584	I*I***
14	1	.02040156	.811	96.395	I*I**
15	1	.01978230	.787	97.182	I*I*
16	2	.01551588	.617	97.799	I*I*
17	2	.01120289	.446	98.245	I*I*
18	1	.01073629	.427	98.672	I*I*
19	3	.00867094	.345	99.017	I*I*
20	2	.00752459	.299	99.316	I*I*
21	3	.00529968	.211	99.527	I*I
22	2	.00434115	.173	99.699	I*I
23	3	.00397999	.158	99.858	I*I
24	2	.00302336	.120	99.978	I*I
25	2	.00055484	.022	100.000	I*I

TITRE: ANALYSE 2

LES VALEURS PROPRES

NUM	ITER	VAL PROPRE	POURCENT	CUMUL	I*I HISTOGRAMME DES VALEURS PROPRES DE LA MATRICE
2	1	.78205085	33.582	33.582	I*I**********I**********I**********I**********
3	1	.56408441	24.222	57.804	I*I**********I**********I**********
4	1	.32317698	13.877	71.681	I*I**********I**********
5	1	.17005637	7.302	78.984	I*I**********
6	1	.10211596	4.385	83.369	I*I********
7	2	.08550771	3.672	87.040	I*I********
8	1	.04771845	2.049	89.089	I*I****
9	1	.04367654	1.875	90.965	I*I***
10	2	.03903826	1.676	92.641	I*I***
11	2	.03386631	1.454	94.095	I*I***
12	2	.03028243	1.300	95.396	I*I**
13	1	.02226026	.956	96.352	I*I**
14	2	.01837292	.789	97.141	I*I*
15	1	.01422430	.611	97.751	I*I*
16	2	.01216192	.522	98.274	I*I*
17	2	.00898532	.386	98.659	I*I*
18	1	.00692419	.297	98.957	I*I*
19	3	.00660619	.284	99.240	I*I*
20	4	.00528385	.227	99.467	I*I
21	3	.00485095	.208	99.676	I*I
22	2	.00339138	.146	99.821	I*I
23	2	.00276733	.119	99.940	I*I
24	2	.00109357	.047	99.987	I*I
25	2	.00030019	.013	100.000	I*I

TITRE: ANALYSE 3

LES VALEURS PROPRES

NUM	ITER	VAL PROPRE	POURCENT	CUMUL	I*I HISTOGRAMME DES VALEURS PROPRES DE LA MATRICE
2	1	.74172795	35.622	35.622	I*I**********I**********I**********I**********
3	1	.53875351	25.874	61.496	I*I**********I**********I**********
4	1	.29396510	14.118	75.614	I*I**********I*********
5	1	.13892031	6.672	82.286	I*I**********
6	1	.09748980	4.682	86.968	I*I*********
7	1	.06514922	3.129	90.097	I*I******
8	1	.04028144	1.935	92.031	I*I***
9	1	.02708767	1.301	93.332	I*I**
10	1	.02453398	1.178	94.510	I*I**
11	2	.02282247	1.096	95.606	I*I**
12	1	.01561878	.750	96.357	I*I*
13	1	.01427963	.686	97.042	I*I*
14	2	.01317047	.633	97.675	I*I*
15	1	.01068181	.513	98.188	I*I*
16	1	.00726754	.349	98.538	I*I*
17	2	.00680561	.327	98.865	I*I*
18	3	.00640753	.308	99.173	I*I*
19	2	.00463823	.223	99.396	I*I
20	3	.00444808	.214	99.609	I*I
21	3	.00285016	.137	99.746	I*I
22	4	.00243416	.117	99.863	I*I
23	2	.00185640	.089	99.952	I*I
24	2	.00097920	.047	99.999	I*I
25	1	.00022018	.011	100.000	I*I

Tableaux E.3

TITRE: ANALYSE 1

I	JI	I	CLT	POID	INR	1=F	COR	CTR	2=F	COR	CTR	3=F	COR	CTR	4=F	COR	CTR	5=F	COR	CTR	6=F	COR	CTR	7=F	COR	CTR
1	C01		999	27	19	53	2	0	-51	1	0	343	66	91	-71	3	1	4	0	0	-601	203	104	482	130	118
2	C02	1	000	397	311	123	77	7	-67	23	3	279	393	91	-65	22	10	2	0	0	-283	407	337	2	0	0
3	C03	1	000	190	321	125	37	4	-69	12	2	402	387	91	-76	14	7	-1	0	0	304	221	185	-307	227	338
4	C04		994	121	291	123	25	2	-65	7	1	336	190	40	26	1	0	-2	0	0	518	450	341	349	204	275
5	C05	1	000	131	691	115	10	2	-39	1	0	-799	486	249	786	469	467	0	0	0	36	1	2	-28	1	2
6	C06		991	38	331	116	6	1	-44	1	0	-1148	607	148	100	5	2	3	0	0	-128	8	7	-108	5	8
7	C07		999	20	521	118	2	0	-59	1	0	-1622	406	157	-1757	476	358	13	0	0	218	7	10	27	0	0
8	C08		719	8	111	112	4	0	-46	1	0	-1026	298	24	-648	119	19	8	0	0	14	0	0	126	4	2
9	C09		962	14	251	114	3	0	-50	1	0	-1318	394	73	-928	196	70	11	0	0	57	1	0	55	1	1
10	C10		989	21	261	110	4	0	-38	0	0	-1360	587	113	-726	168	63	7	0	0	30	0	0	-38	0	1
11	C11		321	3	31	75	2	0	13	0	0	-565	114	31	334	40	21	-1	0	0	-38	1	0	219	17	2
12	C12		218	2	51	5	0	0	88	2	0	-91	2	0	215	10	1	-28	0	0	250	14	2	698	106	23
13	C13		272	2	61	-40	0	0	142	4	0	204	7	0	6	0	0	-24	0	0	570	57	8	887	138	36
14	C14		782	2	81	-426	16	0	1005	86	3	-125	1	0	-199	3	0	-105	1	0	356	11	2	1306	145	58
15	C15		921	2	171	-802	35	2	1992	218	16	-222	3	0	94	0	0	-138	1	0	288	5	2	1603	141	112
16	C16	1	000	10	2251	-2535	118	82	6899	876	837	89	0	0	-22	0	0	181	1	2	-20	0	0	-213	1	9
17	C17		912	1	181	-3259	182	10	4042	280	22	32	0	0	-42	0	0	-686	8	3	-79	0	0	481	4	3
18	C18		692	0	101	-5262	401	13	1773	45	2	61	0	0	-34	0	0	55	0	0	-5	0	0	831	10	5
19	C19		842	1	131	-5048	398	16	1923	58	3	34	0	0	-47	0	0	-499	4	1	-95	0	0	702	8	5
20	C20	1	000	3	1411	-8498	707	308	-2233	49	29	-28	0	0	-27	0	0	-4968	242	626	-12	0	0	-128	0	1
21	C21	1	000	5	2001	-9243	809	500	-3064	89	76	2	0	0	28	0	0	3240	99	365	33	0	0	-38	0	0
22	C22		959	0	171	-7806	567	29	-2007	37	3	22	0	0	18	0	0	591	3	1	-14	0	0	132	0	0
23	C23		662	0	51	-7307	520	7	-2189	47	1	6	0	0	40	0	0	1005	10	1	-5	0	0	-41	0	0
24	C24		812	0	71	-6826	678	14	-2168	68	2	4	0	0	44	0	0	-140	0	0	-26	0	0	-31	0	0
25	C25		648	0	11	-1497	476	1	-525	59	0	1	0	0	134	4	0	179	7	0	-58	1	0	110	3	0
I		I	1933242.	1000			1000			1000			1000			1000			1000			1000			1000	

TITRE: ANALYSE 2

I	JI	I	CLT	POID	INR	1=F	COR	CTR	2=F	COR	CTR	3=F	COR	CTR	4=F	COR	CTR	5=F	COR	CTR	6=F	COR	CTR	7=F	COR	CTR
1	C01		996	13	11	3	0	0	-54	1	0	304	45	4	-129	8	1	2	0	0	-479	113	35	298	44	24
2	C02	1	000	407	311	-112	73	7	-59	20	3	255	370	82	-58	20	8	2	0	0	-289	476	399	56	18	27
3	C03	1	000	196	321	-122	40	4	-67	12	2	393	409	93	-108	31	14	-2	0	0	242	156	135	-352	330	511
4	C04		997	127	301	-119	27	2	-62	7	1	333	204	44	3	0	0	-1	0	0	517	493	397	356	233	337
5	C05	1	000	124	681	-120	11	2	-40	1	0	-737	427	209	819	525	489	-1	0	0	53	2	4	-30	1	2
6	C06		996	41	331	-122	8	1	-48	1	0	-1104	653	154	185	18	8	0	0	0	-87	4	4	-131	9	15
7	C07		998	18	581	-123	2	0	-64	1	0	-1809	448	185	-1790	438	345	9	0	0	192	5	8	-101	1	4
8	C08		684	8	101	-116	5	0	-56	1	0	-1020	363	27	-671	157	22	4	0	0	36	0	0	227	18	9
9	C09		983	11	271	-123	3	0	-62	1	0	-1385	330	64	-988	168	62	9	0	0	94	2	1	392	26	35
10	C10		989	24	281	-117	5	0	-46	1	0	-1350	654	133	-587	124	48	6	0	0	-6	0	0	-26	0	0
11	C11		382	2	41	-68	2	0	-5	0	0	-701	141	4	222	14	1	0	0	0	-26	0	0	-58	1	0
12	C12		402	3	41	-19	0	0	63	1	0	-94	2	0	237	14	1	-7	0	0	204	10	1	355	31	7
13	C13		393	2	51	1	0	0	62	1	0	221	10	0	-37	0	0	-14	0	0	653	83	11	501	49	12
14	C14		738	2	61	336	15	0	876	104	2	-72	1	0	-253	9	1	-33	0	0	374	19	3	216	6	2
15	C15		931	2	161	874	42	2	2247	280	18	-220	3	0	31	0	0	-3	0	0	308	5	2	578	19	14
16	C16	1	000	10	2291	2573	129	88	6604	865	817	51	0	0	-16	0	0	225	1	5	-11	0	0	-39	0	0
17	C17		970	1	221	3300	171	11	4544	323	29	25	0	0	-45	0	0	-814	10	5	-89	0	0	-232	1	1
18	C18		788	0	91	5631	485	12	1550	37	1	43	0	0	-46	0	0	386	2	0	-14	0	0	-31	0	0
19	C19		881	1	131	5100	466	19	2346	99	5	23	0	0	-69	0	0	-1405	35	11	-78	0	0	-49	0	0
20	C20	1	000	3	1231	8243	709	259	-2092	46	23	-49	0	0	-14	0	0	-4803	241	673	-5	0	0	8	0	0
21	C21	1	000	5	2141	9051	835	533	-3162	102	90	-21	0	0	33	0	0	2458	62	301	31	0	0	-8	0	0
22	C22		911	0	161	8088	669	31	-2082	44	3	-6	0	0	18	0	0	778	6	2	29	0	0	-43	0	0
23	C23		766	0	61	8090	630	11	-2502	60	1	-18	0	0	23	0	0	339	1	0	32	0	0	-46	0	0
24	C24		934	0	71	7593	805	16	-2532	90	2	-14	0	0	21	0	0	907	12	2	10	0	0	-30	0	0
25	C25		782	0	11	2378	609	2	-790	67	0	6	0	0	3	0	0	229	6	0	-49	0	0	26	0	0
I		I	1363289.	1000			1000			1000			1000			1000			1000			1000			1000	

TITRE: ANALYSE 3

I	JI	I	CLT	POID	INR	1=F	COR	CTR	2=F	COR	CTR	3=F	COR	CTR	4=F	COR	CTR	5=F	COR	CTR	6=F	COR	CTR	7=F	COR	CTR
1	C01		808	71	11	46	6	0	-27	2	0	294	260	21	-119	43	7	2	0	0	-348	366	132	3	0	0
2	C02		954	299	221	53	18	1	-26	5	0	262	446	70	-115	87	29	1	0	0	-242	382	270	4	0	0
3	C03		993	266	231	60	20	1	-29	5	0	321	569	94	-70	28	10	0	0	0	149	122	90	-3	0	0
4	C04		996	143	221	60	11	1	-29	3	0	184	107	16	149	71	23	0	0	0	447	632	438	-5	0	0
5	C05		997	114	571	60	3	1	-20	0	0	-702	471	191	720	494	424	0	0	0	-64	4	7	-1	0	0
6	C06		991	44	341	59	2	0	-20	0	0	-1110	761	183	349	75	38	0	0	0	-188	22	24	0	0	0
7	C07		984	18	441	58	1	0	-24	0	0	-1694	559	174	-1306	333	219	5	0	0	218	9	13	-2	0	0
8	C08		951	9	191	55	1	0	-24	0	0	-1365	404	56	-1201	313	91	3	0	0	311	21	13	-6	0	0
9	C09		963	8	171	56	1	0	-24	0	0	-1400	429	53	-963	203	53	3	0	0	143	4	3	-12	0	0
10	C10		991	18	321	53	1	0	-16	0	0	-1484	607	139	-877	212	103	5	0	0	32	0	0	0	0	0
11	C11		472	2	31	26	0	0	18	0	0	-763	182	4	269	23	1	4	0	0	-129	5	1	43	1	0
12	C12		805	2	41	-60	1	0	93	2	0	-206	11	0	391	38	2	-4	0	0	135	4	1	41	0	0
13	C13		813	1	41	-101	2	0	114	2	0	130	3	0	136	3	0	-15	0	0	570	52	7	71	1	0
14	C14		741	0	41	-797	33	0	1653	140	2	-79	0	0	-184	2	0	-45	0	0	397	8	1	751	29	6
15	C15		982	0	151	-1968	59	3	4426	297	16	-267	1	0	-3	0	0	8	0	0	297	1	1	2424	89	71
16	C16	1	000	3	2471	-5483	148	103	13127	846	809	56	0	0	2	0	0	463	1	6	-4	0	0	751	3	35
17	C17		977	0	261	-6692	169	12	8953	303	30	22	0	0	-36	0	0	-1734	11	6	-127	0	0	5555	464	620
18	C18		780	0	91	5555	492	13	2836	31	1	46	0	0	-32	0	0	807	2	0	-7	0	0	-4969	94	44
19	C19		890	0	141	5555	467	19	4728	102	6	26	0	0	-66	0	0	-2754	35	11	-97	0	0	-7371	249	183
20	C20	1	000	1	1291	5555	698	253	-4336	50	25	-55	0	0	-13	0	0	-9637	247	680	-1	0	0	1226	4	27
21	C21	1	000	1	2291	5555	826	532	-6566	112	99	-13	0	0	34	0	0	4790	60	292	27	0	0	273	0	2
22	C22		941	0	181	5555	663	34	-4488	51	4	-5	0	0	25	0	0	1821	8	3	36	0	0	-2015	10	10
23	C23		794	0	61	5555	640	11	-5253	68	2	-13	0	0	25	0	0	537	1	0	38	0	0	-241	0	0
24	C24		943	0	71	5555	801	15	-5224	97	3	-13	0	0	17	0	0	1730	11	2	7	0	0	-1062	3	1
25	C25		828	0	11	-4957	653	2	-1653	73	0	0	0	0	-16	0	0	527	7	0	-54	0	0	-427	5	0
I		I	1159113	2d	1000		1000			1000			1000			1000			1000			1000			1000	

Tableaux E.4

Le second facteur distingue trois zones :

. La première zone Z.2.1. est presque identique à Z.1.1.

Les deux autres résultent de la partition de Z.1.2. en deux zones :

. Z.2.2. de ∿ 1800 à ∿ 4500 Hz.

. Z.2.3. de ∿ 4500 à ∿ 7500 Hz.

Le troisième facteur partage Z.1.1. en deux zones :

. Z.3.1. de zéro à ∿ 460 Hz.

. Z.3.2. de ∿ 460 à ∿ 1800 Hz. La partie hautes fréquences du spectre ne contribuant pas à ce facteur.

Le quatrième facteur découpe Z.2.1. en trois zones :

. Z.4.1. de zéro à ∿ 320 Hz

. Z.4.2. de ∿ 320 à ∿ 720 Hz

. Z.4.3. de ∿ 720 à ∿ 1440 Hz . La première et la troisième des zones s'opposant à la seconde.

Le cinquième facteur permet de partager la bande des très hautes fréquences en deux zones :

. Z.5.1. de ∿ 4200 à ∿ 5300 Hz

Et Z.5.2. de ∿ 5300 à ∿ 7500 Hz. La partie basses fréquences du spectre ne contribuant pas à ce facteur.

Le sixième facteur découpe la zone Z.3.1. en deux :

. Z.6.1. de zéro à ∿ 180 Hz

Et Z.6.2. de ∿ 180 Hz à ∿ 460 Hz.

Une autre zone de moindre importance qui apparaît est celle des fréquences moyennes.

. Z.6.3. de ∿ 1580 à ∿ 2500 Hz.

Le septième facteur qui est, rappelons-le, moins stable que les six premiers permet de découper Z.6.2. en deux zones que représentent les canaux trois et quatre et de rappeler la zone Z.6.3. du sixième facteur.

L'analyse factorielle des correspondances permet ainsi de découvrir les interactions attractives ou répulsives entre des zones (quatre pour les moyennes et hautes fréquences, quatre ou six pour les basses fréquences).

.../...

Nous donnerons plus tard une interprétation phonétique des zones et des facteurs ainsi décrits.

Mais de quoi dépendent ces zones (en nombre et en largeur) ?. Du langage, du locuteur ...?. Sont-elles suffisantes pour une description, une caractérisation et une identification du spectre ?. Nous nous proposons de trouver des réponses à toutes ces questions dans nos recherches ultérieures.

E.3.6.3. Avantages de l'enveloppe spectrale

Pour calculer l'énergie du spectre dans un canal, nous pouvons soit faire la somme des composantes du spectre, soit calculer la surface comprise entre l'enveloppe spectrale approchée, les verticales aux bornes du canal et l'axe des fréquences.

Plusieurs analyses ont été faites avec des configurations différentes (nombre et frontières) des canaux et ceci avec les deux procédures du calcul des énergies dans les canaux.

Avec la première procédure, les résultats s'avèrent très sensibles aux variations des frontières des canaux et donc forcément aux variations intra-locuteur et inter-locuteurs.

Alors qu'avec l'utilisation de l'enveloppe spectrale et même pour un choix presque arbitraire des frontières des canaux (par exemple : treize canaux de même largeur entre 200 et 2000 Hz), nous observons une excellente stabilité des six premiers facteurs par leur interprétation et par les zones de fréquence qu'ils déterminent.

Nous estimons que c'est un important critère de stabilité des résultats pour des variations intra-locuteur. Reste à juger cette stabilité pour les variations inter-locuteurs.

E.3.6.4. Interprétation des facteurs

Facteur 1

Le premier facteur est manifestement celui de la fréquence.
Les canaux de la zone Z.1.2. des hautes fréquences avec une contribution écrasante (surtout les C16, C20 et C21 "résidences principales" de l'énergie des Ch et des S)

s'opposent aux canaux de la bande Z.1.1.. Nous observons même un classement fréquentiel quasi-parfait des canaux sur cet axe.

Et d'après le principe des interactions attractives ou répulsives entre les individus et les variables, ce premier facteur fournit aussi une distribution fréquentielle des sons. En descendant la bande passante nous observons les S, Ꞩ, (s emphatique ص), X (ch), T, K, F, = (t emphatique ط), > (h guttural fortement aspiré ح), < (ch allemand خ), J, Q (k prononcé de l'extrême fond de la gorge ق), Z et C (th anglais de thing ث) voir graphique E.5 .

De l'autre côté de l'axe, nous trouvons un paquet regroupé de tous les autres sons.

Les sons suivants S, Ꞩ, F et T sont fortement corrélés ou ont leur plus grande corrélation avec ce facteur et donc s'y expriment le mieux.

Facteur 2

Ce facteur permet de partager le groupe de sons de hautes fréquences qui déterminent le premier facteur. De ce groupe se détachent les S, Ꞩ, Z et C qui s'opposent à tout le reste des sons.

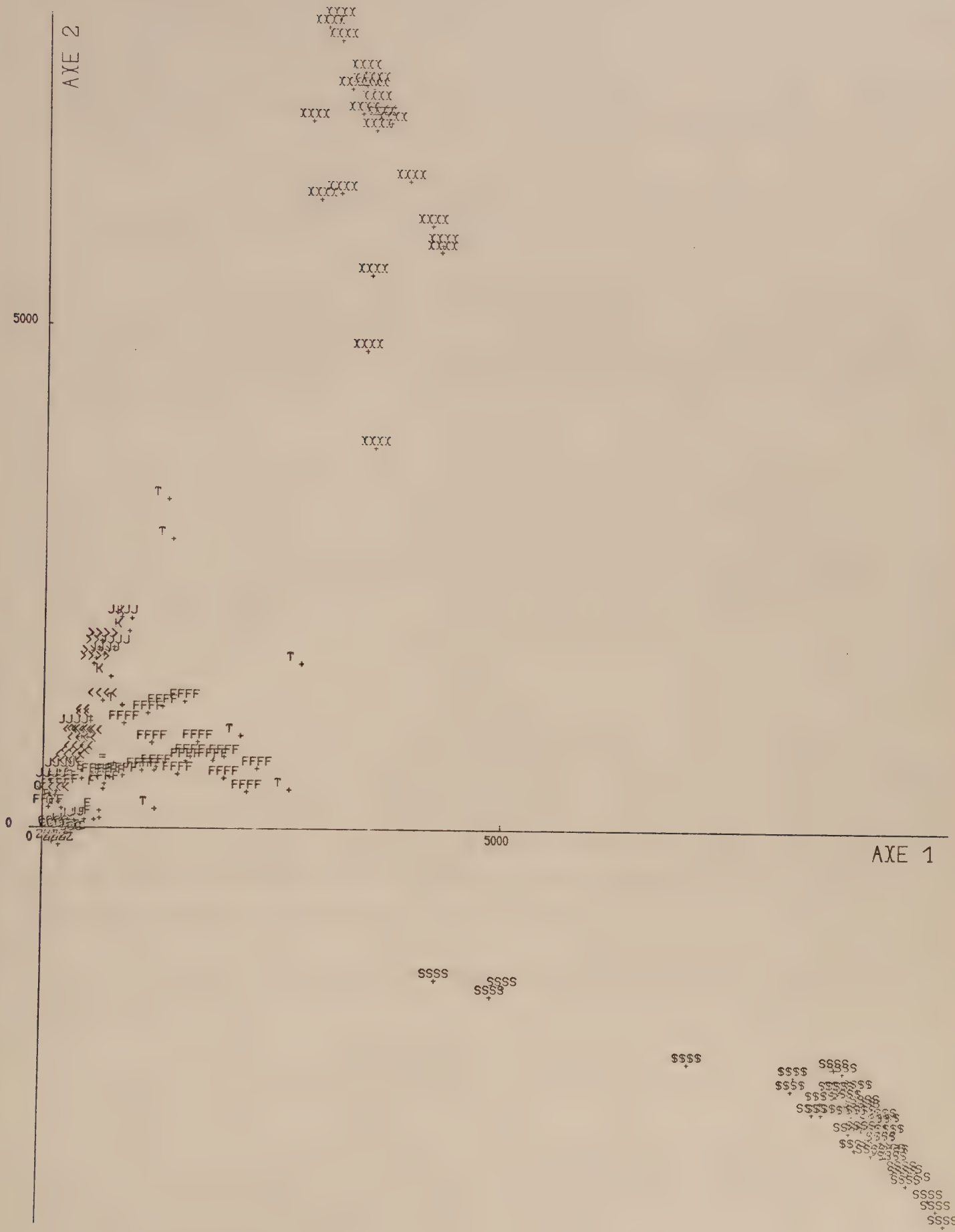
Mais ce sont principalement les X, S, Ꞩ et J ayant des très grandes contributions qui déterminent cet axe.

Les X et les J ont avec ce facteur leur plus forte corrélation.

Nous pouvons ainsi avec le plan factoriel des axes 1 et 2 (graphique E.5) avoir une répartition des sons de haute fréquence qui sera complétée par le cinquième facteur. Les autres sont bien regroupés à côté du centre de gravité dans une petite fenêtre sur ce plan.

Facteur 3

Les canaux C01 à C11 couvrant la bande Z.1.1. de 0 à 1800 Hz environ, contribuent seuls à former ce facteur qui partage cette bande en deux zones Z.3.1. et Z.3.2. dont l'opposition entraîne la formation de deux groupes de sons de fortes



Graphique E.5

Projection des sons de plus grande contribution au plan (1,2)

contributions à ce facteur qui lui donnent la coloration linguistique ou phonétique suivante :

Du côté des canaux C01 à C04 se distinguent par leurs fortes contributions et corrélations les L, M, N, I, Y, U et W.

De l'autre côté du centre de gravité du nuage et avec les canaux C05 à C11, ont des fortes contributions et corrélations les A , ! (e long comme dans كان) , a (a comme dans طار),((r fortement grasseyé غ) et surtout le) (contraction gutturale sonore ع) .

Ce facteur permet aussi un classement d'autres sons. Ainsi les B, D, & (d emphatique ض) et R (roulé) ont tous de fortes corrélations mais aussi les H (aspiré), C, J et Z vont avec le premier groupe. Avec le second groupe, nous observons le F, < (ح), > (خ) , = (ط), T, K et Q (ق) ce dernier ayant avec ce facteur sa plus forte corrélation et occupant sur l'axe une position extrême (voir graphique E.6).

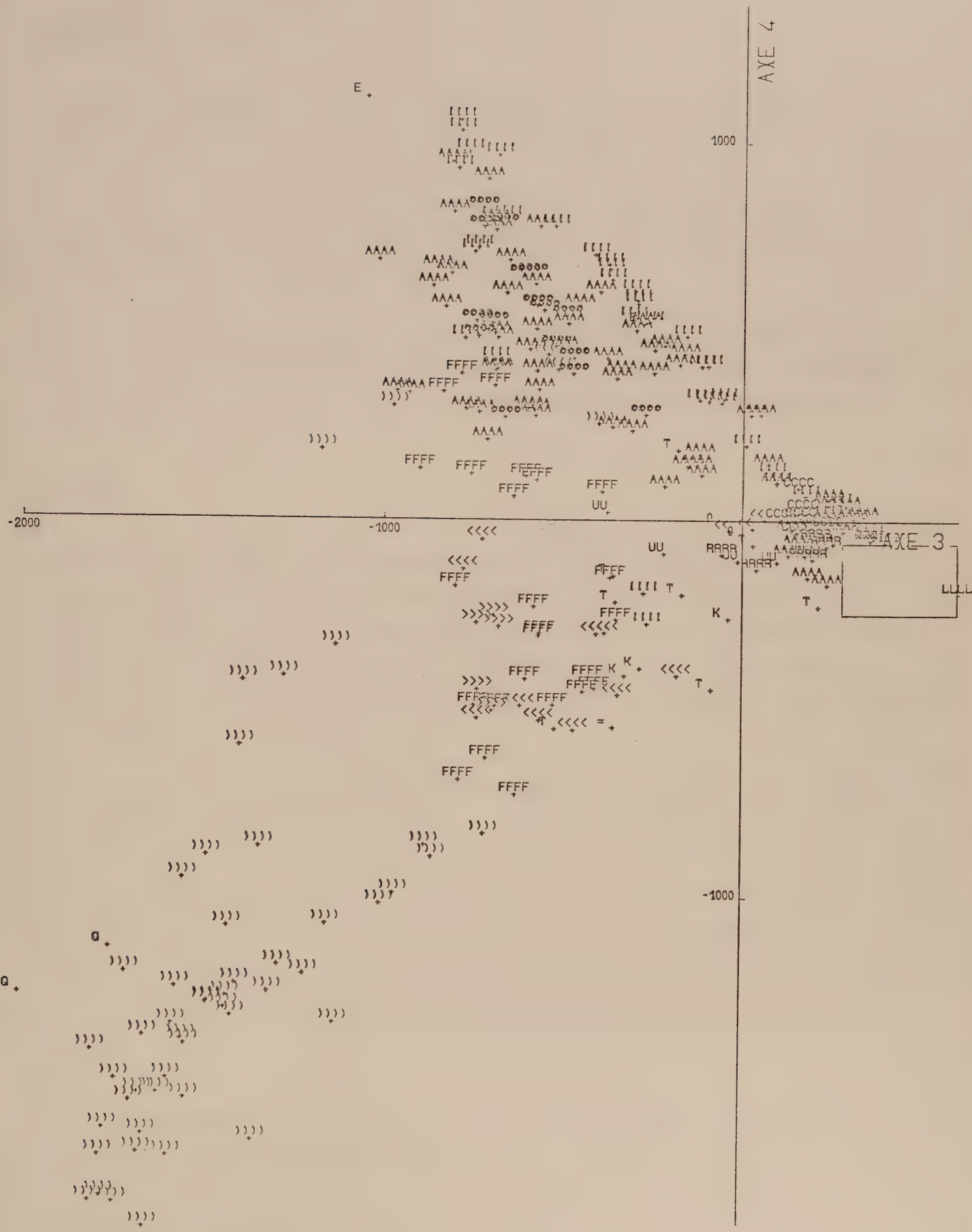
Facteur 4

Ce facteur partage la bande Z.1.1. en quatre zones dont les deux principales couvrent Z.3.2. (qui totalise presque les 4/5 de l'inertie du nuage dans la direction du troisième axe factoriel). Les canaux C05 et C07 caractérisant ces deux zones principales recueillent, au moins, les 2/3 de l'inertie du nuage le long de cet axe et déterminent presque seuls le facteur, sans minimiser les rôles des canaux C01, C02 et C03 d'un côté, C11 et C12 de l'autre.

Sur cet axe, cinq sons émergent avec de très fortes contributions et corrélations.

Le) (ع) pôle principal sur cet axe s'oppose aux A , ! (e long), a (a), E (attaque de voyelle ء) et ((غ).

Bien d'autres sons se trouvent parfaitement rangés sur cet axe. Ainsi du côté du), nous trouvons : le Q (ق) le plus corrélé avec ce facteur et en position très éloignée du centre de gravité mais aussi : les B, T, = (ط) , J, >(خ),<(ح),D, &(ض),R, Z, X, L, M, N, H, I, Y, U et W. Près de l'origine mais de l'autre côté se trouvent les S, $ (ص).



Graphique E.6 : Projection de tous les sons à l'exception de S, Ş et X sur le plan (3,4). Les points à l'intérieur du rectangle étant très nombreux ne sont pas tracés

Facteur 5

Le cinquième facteur est celui des très hautes fréquences. Les canaux C20 et C21 polarisent la presque totalité de l'inertie du nuage le long de cet axe et opposent les S aux $ (s emphatique ص). Nous insistons ici sur le fait qu'il faut considérer ces sons comme ils sont prononcés et non exactement comme ils devraient l'être, surtout que, dans certains contextes phonétiques, un S est pratiquement souvent prononcé comme un s emphatique (comme dans سطح).

Bien que cet axe soit dominé par les S, quelques sons comme les C, F, K, J, Z ... ont des ordonnées stables sur cet axe et sont classés d'une manière bien significative pour toute tentative de segmentation du signal de parole et sa reconnaissance.

Facteur 6

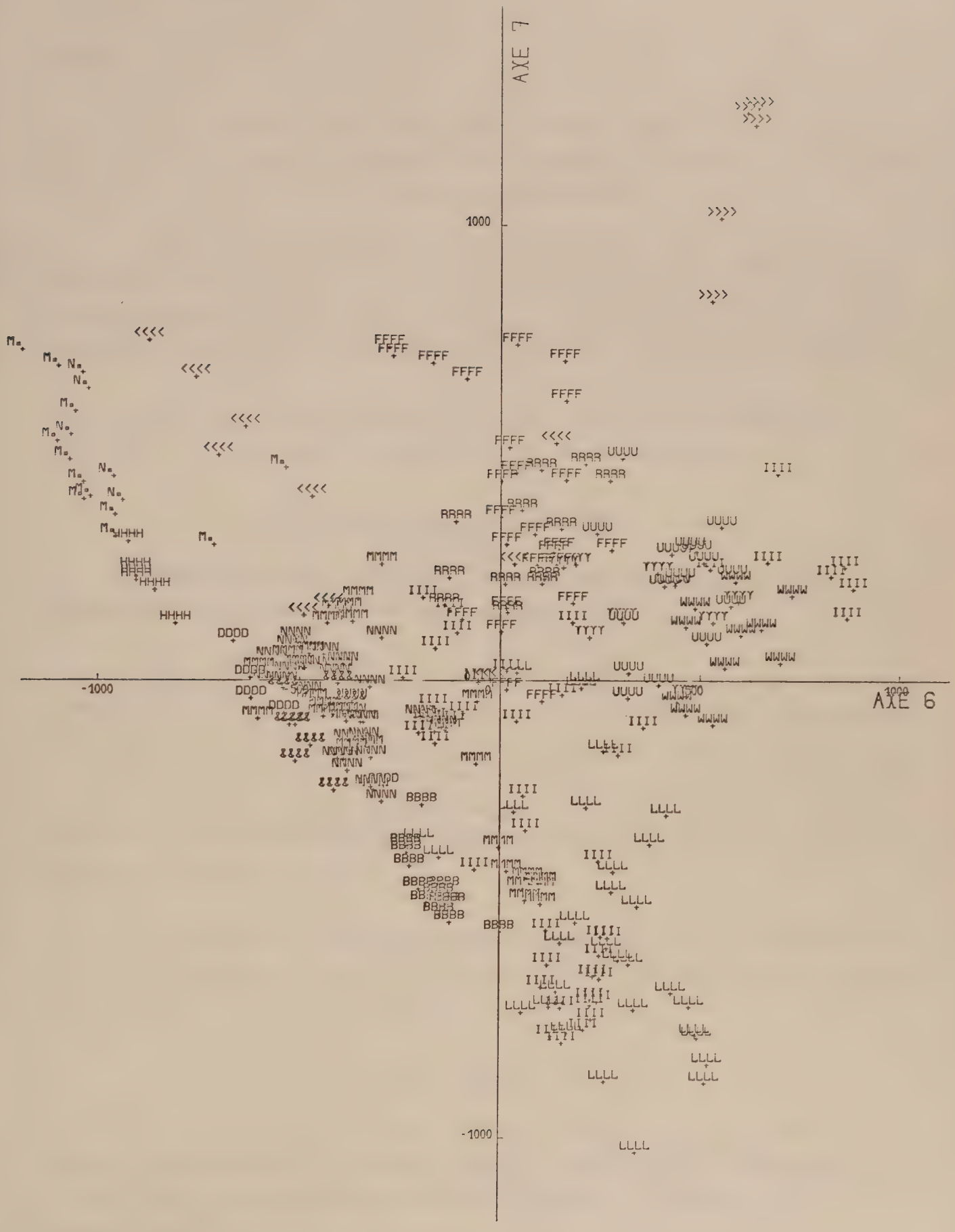
Le sixième facteur est principalement un axe des basses fréquences.
En effet, les canaux C01 et C02 en s'opposant aux canaux C03 et C04 tout en recueillant tous les quatre l'essentiel de l'inertie dans la direction de cet axe, donnent la couleur de ce facteur. Mais une autre zone, celle des canaux C12, ... C15 des fréquences moyennes a une contribution significative et explique les positions de certains sons.

Des quatre premiers canaux C02 et C04 ont les plus grandes contributions et de très fortes corrélations.

En parcourant l'axe six, nous observons les sons suivants qui marquent ce facteur : les L, W, U, I et Y d'un côté, les H, M, N de l'autre. Avec ce second groupe vont les B, D, et & (d emphatique ض) mais surtout les fins des M, N même L qui ressemblent beaucoup au H. Notons que le passage d'un groupe à l'autre n'est pas brutal et que ces deux groupes s'entremêlent légèrement autour du centre de gravité.

Ce facteur permet aussi de classer certains autres sons comme les > (ڎ), <(ڍ) et les B. (Voir graphique E.7.).

.../...



Graphique E.7 : Projection des sons de plus grande contribution au plan (6,7).

Facteur 7

Ce facteur, qui disparaît avec l'utilisation de l'enveloppe spectrale dans la détermination des poids des canaux, est marqué par C03 et C04 mais les deux premiers canaux avec C11 à C15 ont une importance non négligeable.

Bien que des groupes de sons soient bien typés par ce facteur et bien classés comme les B, R, F, >(C) , d'autres groupes de sons avec de notables contributions sont répartis avec des exceptions non négligeables sur cet axe ; nous pouvons ainsi dire que grossièrement les L et I s'opposent aux Y, U et W.

Nous discuterons plus loin de l'opportunité de l'utilisation de ce facteur dans les opérations de segmentation et de reconnaissance.

E.3.7. Analyses de sous-ensembles de sons

Nous avons remarqué (§ E.3.6.) l'importance parfois prédominante que prenaient certains groupes de sons pour certains facteurs.

Bien que nous n'imaginons pas fonder les procédures de segmentation et de reconnaissance de la parole sur les résultats de l'analyse d'un tableau particulier de données, il était intéressant de considérer des tableaux K (I,J) où I est un sous-ensemble de sons, d'analyser ces tableaux et essayer de découvrir des structures plus fines que celles élaborées par l'analyse du tableau original de données et que des propriétés générales assez bien connues (comme l'axe des fréquences) auraient pu masquer.

Pour toutes les analyses des sous-ensembles de sons, nous avons retenu la même configuration de canaux (nombre et bornes) du tableau original établie à partir du signal de parole "complet".

E.3.7.1. Analyse du signal sans X, S, ∅, J

Des 817 spectres de l'analyse 1, nous déduisons, pour les mettre en éléments supplémentaires, les spectres formés partiellement ou entièrement des S, ∅ (ᴖ), X ou J, sons de très grandes contributions aux facteurs 1, 2 et 5 des analyses 1, 2 et 3.

.../...

L'analyse du nouveau tableau ainsi formé permet d'observer que :

- Les quatre premiers nouveaux facteurs correspondent respectivement aux facteurs 3, 4, 6, 7 de l'analyse 1.
- Les nouveaux facteurs 5, 6, 7 malgré l'importance accrue que prennent des sons comme >(ɛ) , <(ɛ̇) , F, ..., n'apportent aucune nouvelle information sur la structure du signal.
- Les spectres X, S, ∅ et J mis en supplémentaires sont moins discernables que dans l'analyse 1.

E.3.7.2. Analyse du signal sans transition

Les transitions reflètent la continuité du fonctionnement mécanique de l'appareil vocal, elles relient les tenues (parties relativement stables du signal) relatives à chaque phonème, et transportent une partie non négligeable de l'information du signal.

Dans un but de reconnaissance spectre par spectre, sans aucune autre segmentation du signal, nous avons éliminé du signal à analyser les spectres des transitions bien que nous sachions qu'en parole continue les phonèmes s'enchaînent rapidement et que la distinction entre tenues et transitions devient parfois difficile et qu'en conséquence de la continuité, l'influence d'un phonème s'étende plus ou moins sur ces voisins. Nous reviendrons sur ce sujet quand nous aborderons les questions des unités phonétiques et de la reconnaissance.

L'analyse des spectres "purs" du signal n'apporte aucun élément significatif nouveau quant aux sept premiers facteurs ou à la répartition des groupes de sons.

E.2.7.3. Analyse des consonnes

L'analyse des consonnes seules (nous avons enlevé tous les spectres du signal de base formés ou contenant des voyelles) avait un double objectif :

La meilleure connaissance de la structure de ces sons.

L'utilisation des résultats de cette analyse pour la reconnaissance de nouveaux spectres si une préalable détection voyelle - consonne était utilisée.

Les cinq premiers facteurs de cette analyse sont les facteurs 1, 2, 3, 5 et 6 de l'analyse 1. Les facteurs 6 et 7 ne sont pas significatifs : en effet, ils sont respectivement dominés par X et)(ڟ où pour chaque facteur des sons d'un même groupe s'y opposent avec de très fortes contributions. Une autre configuration des canaux aurait peut être pu éliminer ce phénomène.

E.3.7.4. Analyses de sous-ensembles de consonnes

Nous avons toujours constaté l'importance que prenaient des consonnes (principalement : X, ʃ, S,), J) dans la détermination d'un ou de plusieurs facteurs. Nous avons, par la suite, analysé l'ensemble des consonnes desquelles sont enlevés et mis en éléments supplémentaires les X, S, ʃ, J ou les) ou tous ces sons ensemble.

Ces nouvelles analyses produisent en partie les structures établies avec un signal "complet" de parole mais permettent aux zones des fréquences moyennes d'avoir une meilleure qualité de représentation et à des sons comme >(ڟ , <(ڟ , F, T, K, Q de se distinguer sur un ou plusieurs facteurs.

Nous tenons à rappeler que pour les analyses précédentes, nous avons pris la même configuration de canaux.

E.3.7.5. Analyses de voyelles

Nous verrons plus loin que le taux de reconnaissance des voyelles est généralement inférieur à celui des consonnes. Nous nous contentons de dire ici que les variations que subissent les voyelles en fonction du contexte phonémique sont plus importantes que celles des consonnes.

Nous citons l'analyse des A, ɑ (a comme dans طار)et ! (e long comme dans كان). Cette analyse ne permet sur aucun des facteurs une réelle discrimination entre les trois groupes. Trois explications sont possibles :

- Notre identification est fausse
- Dans la phonation de l'une de ces voyelles, une ou plusieurs de ces trois formes sont produites.
- Il faut adopter une configuration spéciale des canaux.

Il est difficile de vérifier les deux premières hypothèses ; comme les résultats des analyses avec utilisation de l'enveloppe spectrale sont d'une remarquable stabilité, nous ne pensons pas qu'un changement de configuration de canaux puisse améliorer nettement la discrimination entre ces trois groupes, d'autant plus, vu que les procédures de segmentation et de reconnaissance que nous utilisons ne tirent profit d'aucune information préalable à l'analyse, qu'il paraît impossible d'aligner l'analyse de tout le signal sur une configuration de canaux favorable à un groupe de sons.

E.4. RECONSTITUTION DU SIGNAL DE PAROLE

Tout spectre -tranche du signal de durée fixe- est caractérisé par son profil spectral et par suite par ses coordonnées dans l'espace des premiers facteurs qui totalisent une grande part de l'inertie du nuage et donc de l'information des données.

L'analyse factorielle des correspondances offre -par la formule dite de reconstitution- le moyen pour un individu i (spectre), (comme pour une variable) qu'il soit du tableau principal ou un élément supplémentaire, de restituer approximativement son profil c'est-à-dire $K(i,j) \mid j = 1, \ldots C$; C étant le nombre de canaux mais aussi l'ensemble des 512 valeurs (pour un spectre de 1024 points de durée 50 ms) représentant l'énergie du spectre.

Pour montrer la validité et le bien fondé du codage par facteurs, nous avons entrepris plusieurs expériences ; voici la description rapide de l'une d'elle :

Les variables principales étant au nombre de 25 (canaux), nous ajoutons comme variables supplémentaires les 479 premières valeurs des 512 ; d'ailleurs, nous avons remarqué que nous pouvons dscendre à des fréquences inférieures et réduire ce nombre de variables supplémentaires. Cette expérience pouvant être menée avec les 479 valeurs originales ou déduites à partir de l'enveloppe spectrale approximée (dans ce second cas, l'utilisation de la fréquence fondamentale paraît nécessaire pour placer ultérieurement les valeurs sur l'axe des fréquences).

Les coordonnées, de ces variables supplémentaires, livrées à la formule de reconstitution, permettent d'obtenir des valeurs approximatives des 479 valeurs originales.

.../...

Avec des phases aléatoires construites avec des suites de nombres au hasard, nous appliquons la transformée de Fourier inverse sur les valeurs approximées. Nous procédons ainsi pour tous les spectres successifs.

La conversion numérique / analogique de ces dernières valeurs reconstitue le signal de parole. Bien que le signal reconstitué soit intelligible, surtout pour les sons de moyenne et haute fréquences, il est cependant accompagné d'un bruit régulier résultant du raccordement des spectres reconstitués et nuisible surtout pour les sons de basse fréquence ; d'ailleurs en filtrant les très basses fréquences (0 à 200 ou 300Hz), le confort de l'écoute du signal s'améliore nettement mais ceci détruit une partie non négligeable du signal.

Nous avons essayé, sans grand succès, de résoudre le problème du raccordement des spectres par fusion de parties adjacentes des spectres consécutifs.

Nous nous proposons de résoudre ce problème pour faire du codage par analyse factorielle des correspondances une technique utilisable dans les moyens de transmission du signal.

CHAPITRE F

SEGMENTATION DE LA PAROLE

Les discours peuvent être saisis comme une suite d'éléments de sens et de sons distingués suivant des principes qui sont ceux qui président à l'écriture des mots et des lettres. Cependant la chaîne parlée n'est pas une séquence discrète mais un continuum physique ; et la parole n'existerait pas sans les objets dont elle parle. D'où le problème des unités, aussi bien les unités de son que les unités de sens, que l'analyse statistique traitera.

F.1. LE PROBLEME DES UNITES

L'histoire de l'écriture suffit à révéler la complexité du problème des unités. Pour celui qui pratique dès l'enfance l'écriture, la langue en est venue à s'identifier avec le jeu des lettres : on croit prononcer et l'on tend à prononcer comme on écrit : si parfois une divergence est apparente, on se contente d'une explication historique. Cependant, les systèmes de notation des diverses langues sont d'une très inégale valeur. D'une part les sons ne se présentent pas découpés mais enchaînés, d'autre part, ils ne se répètent jamais identiques à eux-mêmes. Ainsi "la délimitation des unités concrètes d'une langue est un problème si délicat qu'on se demande si elles sont réellement données ... sans qu'on puisse douter cependant qu'elles existent et que c'est leur jeu qui constitue la langue" (De Saussure 1910).

D'autre part, la segmentation est loin d'être hors de conteste, d'autant plus que cette opération peut se présenter différemment selon les langues : dans les langues sémitiques, les consonnes ont une grande importance ; en chinois, la syllabe est une unité très forte, étant souvent à la fois unité de sens et de son.

.../...

De toute façon, il y a dans toute segmentation une part d'arbitraire surtout si l'on se demande comment opèrent les sujets parlant eux-mêmes ; c'est là un problème général de psychologie qui interfère avec la linguistique : quelles sont par exemple les unités perceptives de celui qui écoute, le phonème certainement pas ..., la syllabe, le mot ?. Et, par suite, quelles unités - segments - chercher, quelle analyse faire, phonétique , phonologique ou autre ?.

Nous pensons que pour passer d'un discours à une suite de segments ou de signes, il faut découper et classer ?. Mais il est quasi-inévitable de partir d'un discours qui soit segmenté suivant des règles imparfaites qui ne sont pas le fruit d'amples recherches, ni d'une réflexion linguistique profonde et où l'usage séculaire a pu autoriser des inconséquences manifestes. C'est sur ces unités en usage qu'on travaillera d'abord ; non toutefois pour se vouer sans retour à la première segmentation reçue. L'analyse des données permet au contraire de critiquer cette segmentation et la perfectionner : ici on soudera deux éléments qui ne vont jamais l'un sans l'autre ; ailleurs, on scindera ... .

Avant d'exposer nos travaux sur la segmentation et les principes sur lesquels nous nous basons, nous présentons quelques théories et réflexions linguistiques sur la question des unités.

F.1.1. La double articulation et l'écriture alphabétique

Martinet A. (1965) écrit : "Le langage humain est articulé sur deux plans, celui où, pour employer les termes du parler de tous les jours, les énoncés s'articulent en mots, et celui où les mots s'articulent en sons ... . L'articulation linguistique en unités significatives résulte du fait que pour exprimer une situation qui peut paraître au sujet comme un tout absolument unique, inanalysable et irréductible à ses expériences antérieures et à celles d'autrui, il faut utiliser une succession d'unités dont chacune a une valeur sémantique particulière ... . L'exigence d'économie requiert que les segments minimaux pourvus de sens (mots ...) soient eux-mêmes articulés en sons élémentaires suivant un petit nombre de types sonores dépourvus de sens (les phonèmes) qui, sans cesse reproduits, suffisent à tout dire".

"La théorie de la double articulation aboutit à distinguer nettement parmi les productions vocales entre des faits centraux, ceux qui entrent dans le cadre qu'elle délimite, et des faits marginaux, tous ceux qui, en tout ou en partie,

.../...

échappent à ce cadre. Les faits centraux ainsi dégagés, signes et phonèmes, sont ceux dont le caractère conventionnel, est le plus marqué ; ils sont d'une nature que d'après les mathématiciens, on nomme "discrète", c'est-à-dire qu'ils valent par leur présence ou leur absence, ce qui exclut la variation progressive et continue".

Nous pensons que l'examen des analyses que nous avons faites, peut modifier des conceptions sur la segmentation de la chaîne parlée ; au niveau de la première articulation (place des silences) ; au niveau de la seconde articulation : quelles sont les natures et les oppositions réelles entre les phonèmes ?.

D'autre part, De Saussure F. pense que dans la mesure où la parole, l'exécution a ses normes propres, elle ne s'oppose pas absolument à la langue, mais que dans les faits, il y a de celle-ci à celle-là une gradation que l'on peut rapprocher de l'échelle que Benzécri J.P. construit ainsi :

- La conscience (du sujet parlant)
- L'inconscient (du sujet parlant)
- Le patrimoine actuel (de la communauté linguistique)
- Le patrimoine potentiel (de la communauté linguistique)

F.1.2. Le trait pertinent et l'école de Prague

L'école de Prague a, par les travaux notamment de Trubetskoj N., renouvelé l'étude linguistique des sons, en créant la phonologie. Le discours (ou chaîne parlée) est découpé en une suite de segments, correspondant à des phonèmes (qui coïncideraient avec les lettres). Linguistiquement, un phonème est un élément du système phonologique d'une langue donnée (chaque langue a le sien) ; matériellement, il se réalise comme un ensemble de segments de chaînes parlées. Le système des phonèmes (qui n'existent pas matériellement par eux-mêmes) est un équilibre qui peut évoluer, entraînant un redécoupage, une redistribution des sons en un nouveau système.

L'école de Prague vise à décrire les phonèmes, sans considérer les détails des sons, par un nombre minimum de "traits pertinents" : les caractères qui sont nécessaires et suffisants pour les distinctions que fait la langue. (Ainsi, en

.../...

français, *s* peut être prononcé avec un défaut de langue - l'on devrait dire un défaut de parole ! - sans qu'aucune confusion puisse en résulter, mais en anglais, en arabe il y a deux phonèmes. Ces traits pertinents sont des coordonnées discontinues prenant un petit nombre de valeurs (souvent deux) et caractérisant le phonème (exemple : le trait de "sonorité" est une coordonnée à deux valeurs : sonore (*b*), sourde (*p*) ; "le point d'articulation" prend plusieurs valeurs : labial (*b*), dental (*d*), guttural (*g*) ...).

Quoique la détermination du système phonologique d'une langue suive, en principe, une procédure systématique, le résultat de l'opération n'est pas nécessairement unique : c'est que l'on réunit sous un même phonème non seulement toutes les réalisations individuelles d'un même projet sonore ou "phone", mais des sons relevant de projets distincts (que les principes de la phonologie amènent à regrouper) : les allophones du phonème . Le regroupement des allophones ne s'impose pas toujours, ni même une segmentation de la chaîne.

Parmi les autres théories linguistiques, nous citons celle de l'analyse distributionnelle de Harris qui a influencé Chomsky dans la création de sa linguistique.

Nous nous sommes inspirés, à ce sujet, des oeuvres de Benzécri J.P. et principalement "Pratique de l'analyse des données - Linguistique et Lexicologie" (1981).

F.2. SEGMENTATION AUTOMATIQUE DE LA PAROLE

Cette opération consiste à passer automatiquement du signal sonore continu à une suite d'éléments discrets réalisant une sorte de transcription phonétique dans laquelle on cherchera à reconnaître ultérieurement des mots, des phrases, un discours. La nature d'un segment dépend des critères et du type d'analyse employés.

Sans vouloir citer les méthodes utilisées, nous disons que dans le cas des méthodes fréquentielles, la segmentation interviendra après l'analyse spectrale sur des critères comme l'instabilité du spectre, la variation de son énergie, l'évolution et la place des formants ... . Le voisement est aussi un critère de segmentation

.../...

Dans le cas des méthodes temporelles, la segmentation peut se faire en début d'analyse sur des critères d'énergie ou de variation du taux de passage par zéro.

F.2.1. Segmentation par analyse des correspondances

F.2.1.1. Découpage aveugle

Dans les tableaux de données que nous analysons, l'ensemble I résulte d'un découpage aveugle de la chaîne parlée ; découpage qui n'a aucune chance d'avoir isolé des unités recevables telles que des phonèmes. Du strict point de vue distributionnaliste, le corpus devient muet ; tout texte semble écrit dans une langue morte encore indéchiffrée.

L'étude statistique en cherchant des cooccurences au sein des segments d'une taille fixée, doit dans ses conclusions rompre ce silence. Justement, c'est en regroupant les segments i qui se ressemblent, qu'on constituera des unités proches des phonèmes, dont on observera l'enchaînement.

Observer et analyser la chaîne sonore, c'est suivre les variations des spectres - segments de taille fixée - au cours du temps dans l'espace des variables : tantôt le spectre stationne dans le domaine d'une voyelle ; tantôt d'un tel domaine, il s'approche ou s'éloigne en suivant un processus consonantique. Précisons alors concrètement la pratique de notre segmentation.

F.2.1.2. Segmentation

Notre spectre - segment de la chaîne d'une durée de 50 ms - est un point de l'espace des variables (canaux) de dimension 25.

Après analyse factorielle, le profil spectral est caractérisé par les six ou sept premiers facteurs issus du tableau K(I,J). Nous pensons que cette description synthétique des profils est bien suffisante ; ces facteurs représentent la structure générale du langage étudié où chacun exprime une qualité - parfois difficile à élucider - traduite souvent par l'opposition sur l'axe factoriel entre deux tendances ayant des abscisses opposées ou par un étirement du nuage de points d'une extrêmité de l'axe à l'autre.

.../...

De plus, quelles que soient les difficultés rencontrées dans l'interprétation individuelle des facteurs, il nous paraît à peu près démontré que l'ensemble des sept (ou six) premiers facteurs, fournit une représentation géométrique suffisamment précise des sons instantanés (identifiés aux spectres) par des points dans un espace de faibles dimensions.

Pour suivre l'évolution la plus précise possible du spectre de durée 50 ms et pour éviter qu'une unité phonétique scindée en deux par le premier découpage aveugle soit mal étudiée, et donc mal caractérisée, nous prenons des spectres chevauchant dans le temps, de 37,5 ms (translatés de 12,5 ms). Cette opération s'est révélée très profitable, voire indispensable ; c'est le ralenti d'une image qui permet de saisir les détails du profil spectral qui évolue très rapidement dans le temps.

Notre spectre est un point qui évolue dans l'espace des sept premiers facteurs, il est attiré selon son profil dans la direction d'un axe, et généralement de plusieurs axes par la tendance ou les tendances que représentent ces axes dans un sens ou dans un autre.

Nous observons la trajectoire mobile du spectre et nous disons qu'il y a, d'une manière générale, une coupe à chaque fois que la trajectoire change significativement de sens, en traversant le centre de gravité (pour un seul axe) ou en rebroussant chemin après s'être approchée du centre de gravité dans l'espace de plusieurs facteurs ou des sept facteurs.

Ainsi la segmentation peut être faite à plusieurs niveaux :

- une segmentation "grossière" par un facteur.

- une segmentation plus fine en considérant la trajectoire du spectre dans un sous espace ou dans l'espace des sept premiers facteurs.

Mais nous avons remarqué que deux facteurs différents (par exemple le troisième et le quatrième dans les figures F.1 et F.2) donnent lieu à des coupes aux mêmes endroits (exemples : après les spectres 190, 206, 227 et 260) mais aussi à des endroits distincts mais très voisins (après 240 et 242 pour le passage de ! à <,

.../...

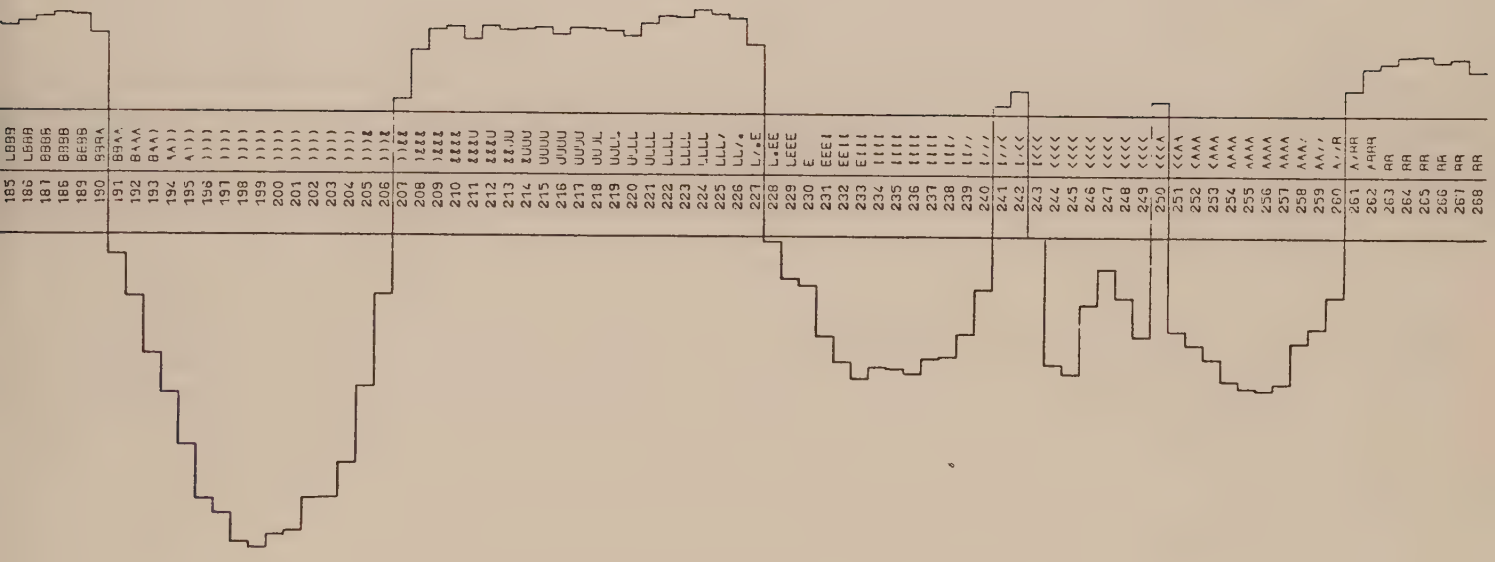


Figure F.1

Tracé de l'abcisse des spectres sur l'axe 3

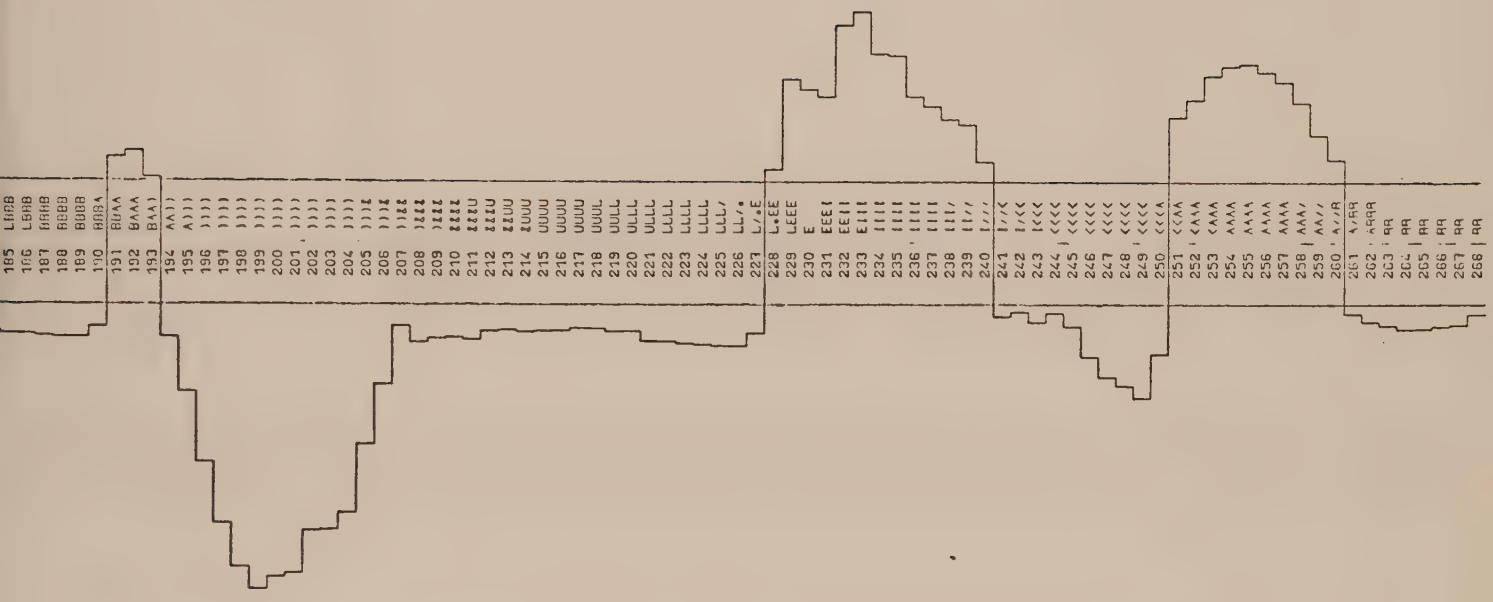


Figure F.2

Tracé de l'abcisse des spectres sur l'axe 4

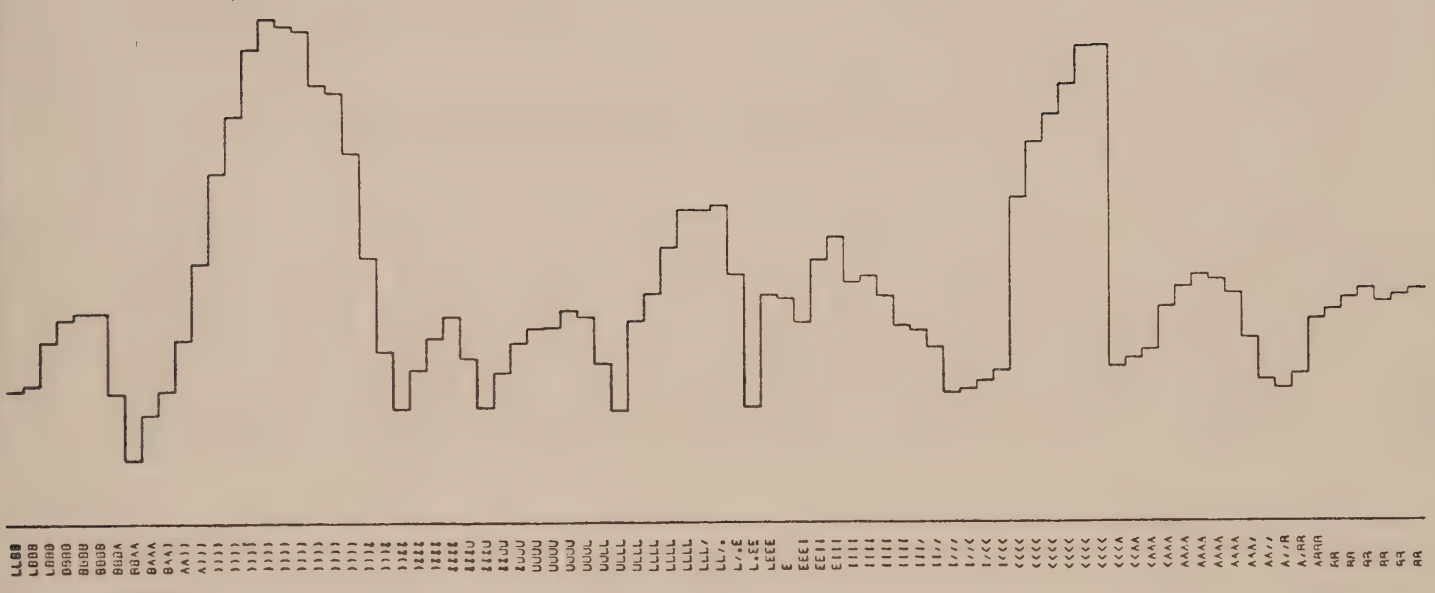


Figure F.3

Tracé de la distance à l'origine des spectres dans l'espace des sept premiers facteurs.

et après 249 et 250 pour le passage de < à A). Ceci est dû au fait que les deux facteurs n'ont pas parfaitement le même degré de sensibilité aux changements de tendance pour le profil spectral. On peut même supposer que les facteurs sont associés aux diverses déformations du conduit vocal, lesquelles ne sont pas exactement synchrones dans l'articulation d'un son.

Tenir compte de toutes les coupes obtenues par tous les facteurs, conduit forcément à un nombre excessif de segments qui peuvent ne pas correspondre aux réalités du signal de parole.

Plusieurs solutions étaient alors envisageables :

- éliminer un segment de faible longueur . Alors ceci conduit à éliminer les sons brefs comme *t*, *k* ... mais nous pensons que ces sons brefs sont bien repérables surtout dans le plan des facteurs 1 et 2 et qu'il serait possible de tester à chaque fois que nous avons un segment court, s'il s'agit des sons brefs, par comparaison aux coordonnées de ces sons brefs, des coordonnées des spectres du segment ou par d'autres moyens.

- conjuguer toutes les coupes en cherchant pour chaque spectre sa distance au centre de gravité du nuage et ceci dans l'espace des sept premiers facteurs. La figure F.3 représente le tracé de cette distance pour un signal d'une durée d'une seconde environ.

Notre algorithme, très simpliste, consiste à décider de l'existence d'une coupe aux creux de cet histogramme donc à chaque fois que la courbe finit de s'approcher de l'axe horizontal pour commencer à s'éloigner. Toutefois, nous négligeons les creux non significatifs (exemple : aux spectres 234 et 266).

En comparant les figures F.1, F.2 et F.3, nous déduisons les observations suivantes :

- L'utilisation de tous les facteurs (Figure F.3) raffine nettement la première segmentation (Figures F.1, F.2) en détectant les passages entre & (ض) et U, entre U et L et entre E et !.

.../...

- Couper aux minima de la distance au centre de gravité conduit à perdre des coupes déjà obtenues avec au moins l'un des facteurs comme le cas du passage de A à)(ع) dans F.2 qui disparaît dans la Figure F.3. Bien que l'existence du A (très court) soit détectée par le facteur 4 (Figure F.2),nous pouvons nous demander s'il n'a pas été étouffé par le)(ع) et si en réalité en écoutant un signal construit avec les segments détectés dans la figure F.3, nous n'obtenons pas un signal identique à l'original !.
N'y-a-t-il pas souvent un certain A dans le début d'un)(ع) ?!.

Notre programme de segmentation est actuellement celui du minimum de la distance du profil spectral du signal au centre de gravité du nuage dans l'espace des sept premiers facteurs issus de l'analyse des correspondances. Comme pour toute procédure de segmentation automatique, deux types d'erreurs sont possibles :

- Omettre de détecter une coupe qui s'impose par le contenu phonétique du signal.

- Décider d'une coupe que la phonétique traditionnelle n'aurait pas laissé prévoir dans certains cas. Mais existe-t-il en réalité une segmentation totalement objective de la chaîne parlée en segments idéaux ?!.

Sans vouloir trancher d'une manière définitive et rigide la question des segments idéaux, nous disons que l'observation et l'analyse des segments ainsi détectés après une analyse objective des correspondances,nous permettent de dire que ces segments doivent être étiquetés par des "phones", (variantes des réalisations des phonèmes) : chaque segment relevant d'un phone unique, le système des phones pouvant être découvert avec classification automatique hiérarchique (C.A.H.) effectuée sur l'ensemble des segments découpés par notre programme (quitte à perfectionner ensuite le système ainsi construit) ou sur l'ensemble des spectres.

Notre programme appliqué à un signal de parole de 25 secondes donne les résultats suivants :

- Il déclare 218 coupes dont 201 correspondent parfaitement au découpage en phones auquel nous nous attendons, 6 coupes moins bien placées que les premières et 11 coupes qui peuvent être considérées comme supplémentaires et dont certaines prouvent qu'un phonème au sens traditionnel du terme peut être considéré comme

résultant de la jonction de deux phones. Nous pouvons être satisfaits de ce degré de réussite.

Par ailleurs, 49 coupes détectables sur un ou plusieurs facteurs ne sont plus détectées par le minimum de la distance au centre de gravité.

Rappelons que déjà une première segmentation, la détection du silence est effectuée, qu'elle fonctionne bien et ne rentre pas dans les statistiques que nous avons données.

Nous avons vu que dans notre programme, nous ne tirons pas totalement profit de toute l'information et de toute la description d'une tranche du signal par les facteurs de l'analyse des correspondances. Nous devons conjuguer les deux procédures pour aboutir à une meilleure réussite dans notre segmentation : c'est une tâche que nous entreprendrons par la suite.

Cependant deux réflexions au sujet des unités s'imposent et découlent de nos nombreuses expériences et analyses : particularité du dernier phonème d'un mot et influence du débit de parole et de l'articulation en matière de segmentation.

F.2.2. Particularité du dernier phonème d'un mot

L'étude du dernier phonème d'un mot paraît intéressante à deux niveaux : la segmentation et la reconnaissance automatiques.

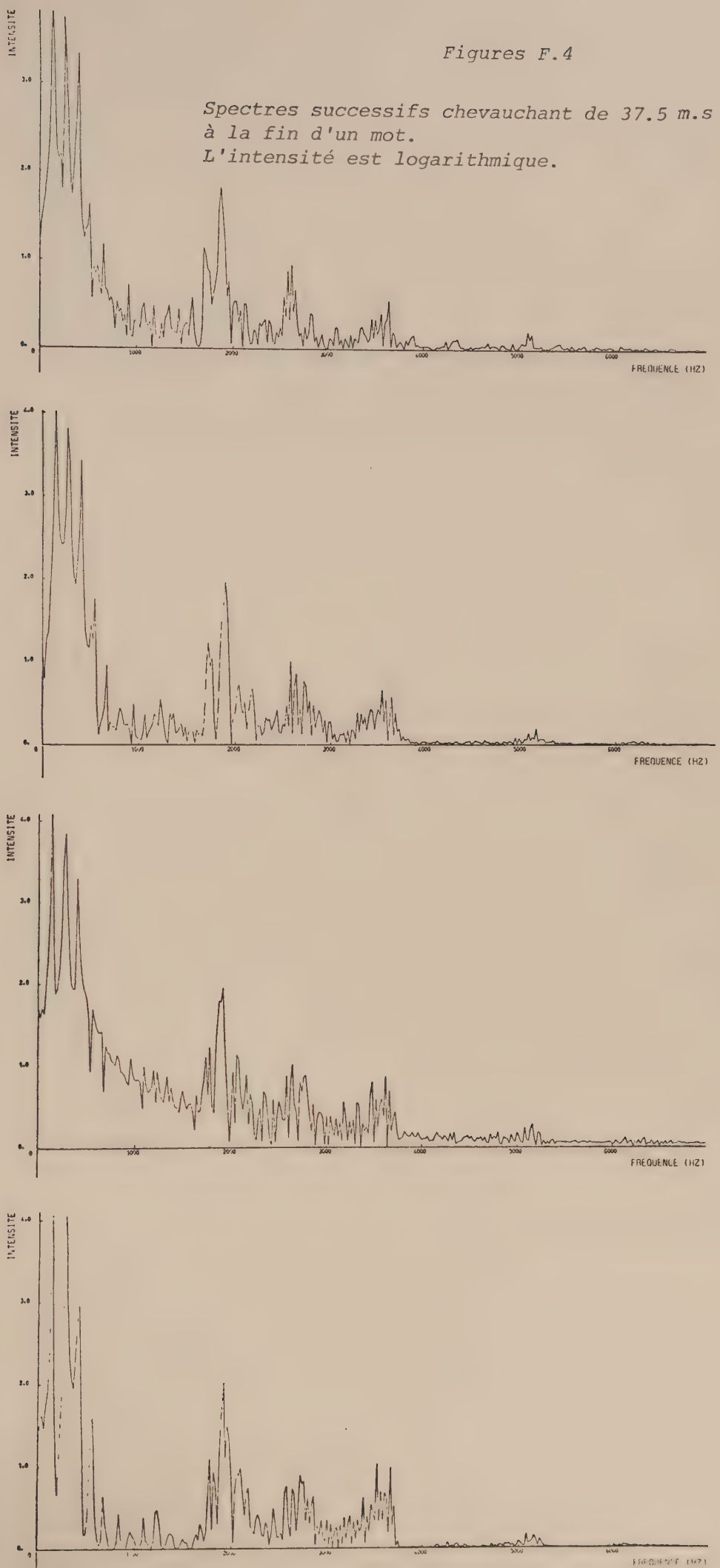
Les Figures F.4 représentent les spectres successifs, chevauchement de 37,5 ms, de la voyelle *i* à la fin du mot (AR&I ارضِ). La simple observation de ces figures montre l'évolution des harmoniques et du second formant. Il n'est pas étonnant que l'analyse factorielle scinde ce phonème en deux et donne ainsi deux segments. En réalité, la coupe peut être faite par les facteurs 5, 6 et 7, ces facteurs ne coupant pas exactement au même endroit, situation que nous avons déjà connue et commentée ; alors que le minimum de la distance au centre de gravité coupe au spectre 72.

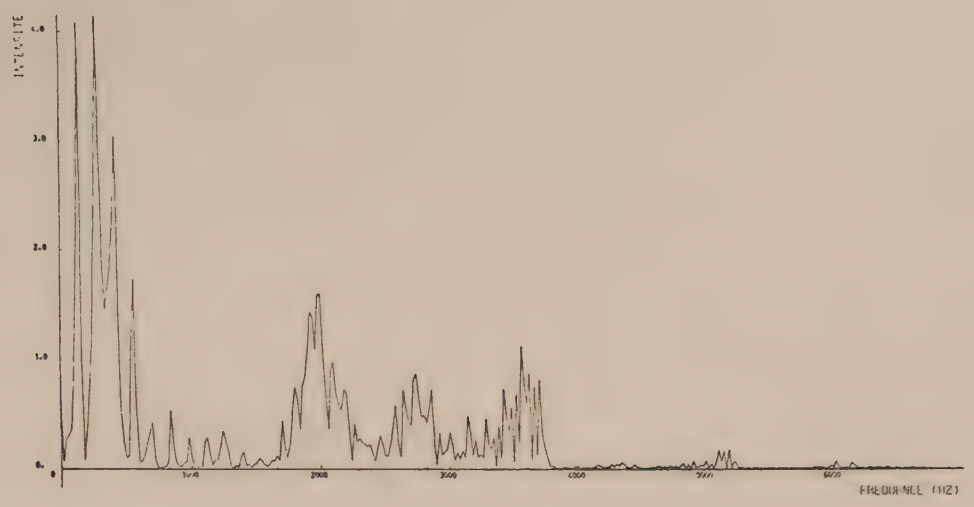
Ce phénomène a été souvent observé avec la voyelle dernier phonème du mot mais aussi avec quelques consonnes comme le *m* et le *n* toujours en fin du mot.

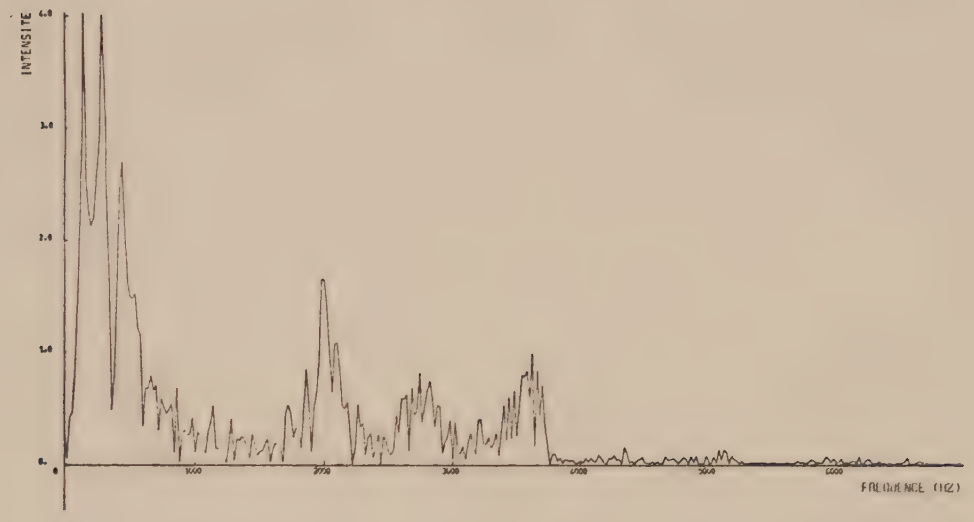
.../...

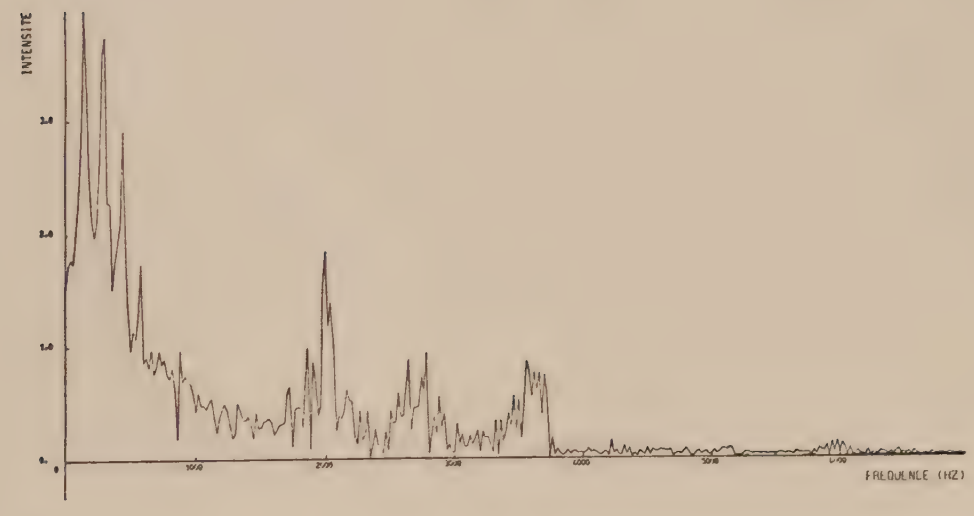
Figures F.4

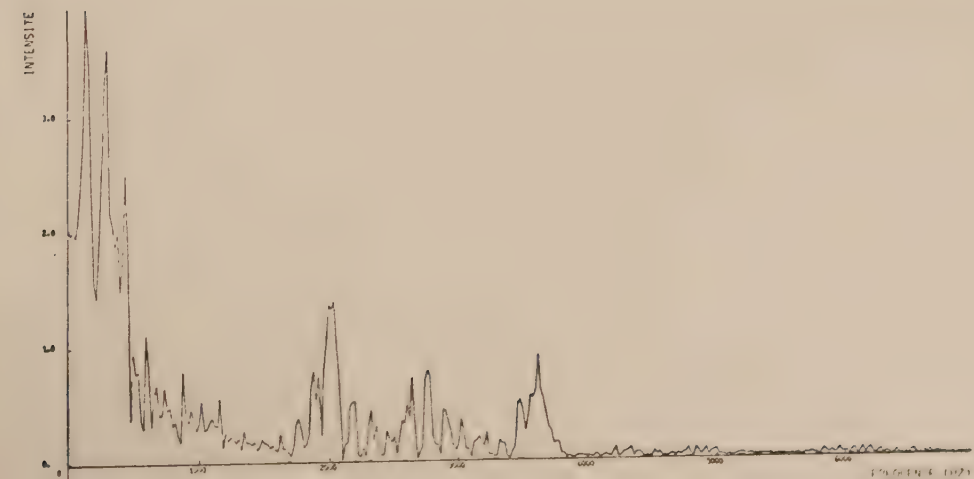
Spectres successifs chevauchant de 37.5 m.s de la voyelle i à la fin d'un mot.
L'intensité est logarithmique.

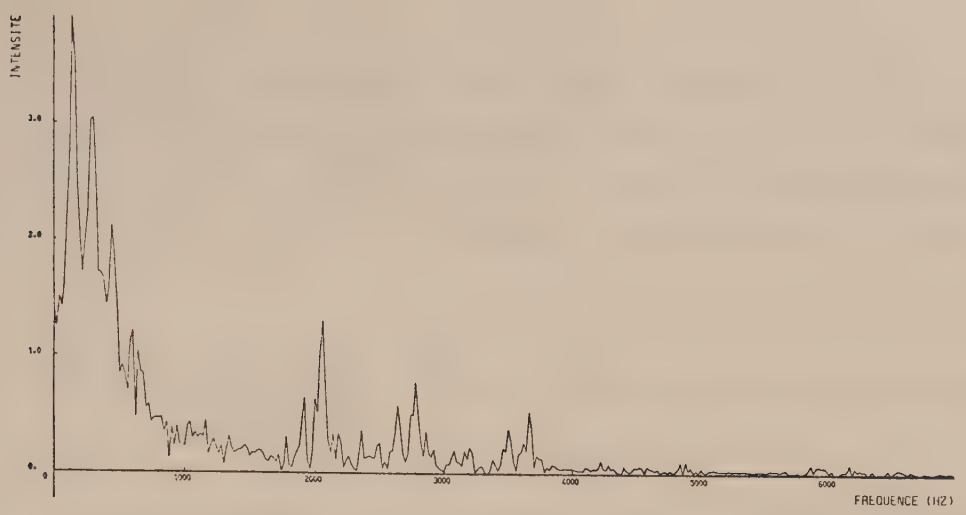



INTENSITE
FREQUENCE (HZ)

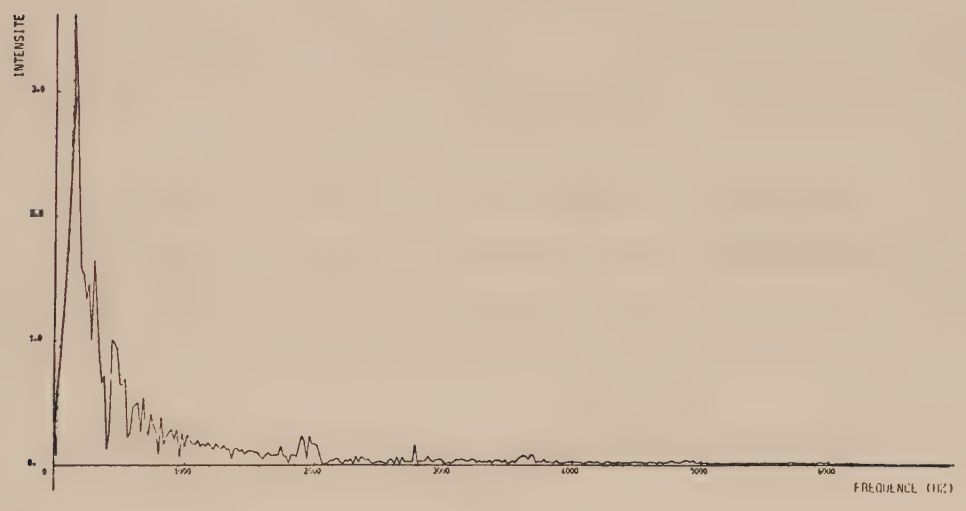

INTENSITE
FREQUENCE (HZ)


INTENSITE
FREQUENCE (HZ)


INTENSITE


INTENSITE
FREQUENCE (HZ)

INTENSITE
FREQUENCE (HZ)


INTENSITE
FREQUENCE (HZ)

INTENSITE
FREQUENCE (HZ)

La prise en considération des seconds segments de ces phonèmes s'est révélée indispensable pour la reconnaissance. Nous pensons que, généralement, quand un phonème "meurt" lentement à la fin du mot en correspondant à la libération du reste de l'énergie vocale, il correspond acoustiquement à des sons comme le *h* aspiré.

F.2.3. Influence du débit de parole et de l'articulation sur la segmentation

Nous avons travaillé principalement sur le signal de parole obtenu par la lecture d'un texte avec un débit normal et une articulation naturelle. Mais comme l'un de nos objectifs est la segmentation et la reconnaissance automatiques, nous étions tentés par l'expérience suivante : pourquoi ne pas lire le même texte avec une accentuation attentive et une articulation, non exagérée, mais qui explicite tous les phonèmes du discours.

La lecture du même texte a duré dans ces conditions, un temps 4/3 de fois plus long.

L'analyse d'une tranche de ce nouveau signal montre que les phonèmes et surtout les voyelles ont une tenue beaucoup moins stable et sont segmentées en plusieurs unités ; nous nous demandons ainsi si la tenue d'un phonème, surtout la voyelle, n'est fonction,non seulement de la consonne qui le précède ou (et) qui la suit mais aussi de la nature de l'articulation et du débit de parole.

Nous nous proposons, dans nos recherches ultérieures, d'étudier quantitativement (durée dans le temps) et qualitativement (segmentation du phonème) la tenue surtout des voyelles en fonction de la vitesse du débit de parole, de la nature de l'articulation et du locuteur.

D'ailleurs l'observation du signal dans le temps permet de se rendre compte des variations dans chaque phonème.

Nos premiers travaux ne nous autorisent pas à dire qu'une telle lecture pourrait améliorer la segmentation et surtout la reconnaissance.

.../...

F.3. LES CONSONNES ET LEUR EXISTENCE PROPRE

Rodet X. écrit dans sa thèse (1977) : "il est admis que les consonnes n'ont pas d'existence propre mais qu'elles apparaissent en association avec les voyelles et qu'elles s'actualisent dans la transition qui précède ou suit la voyelle. Cependant l'examen du signal vocal permet de reconnaître des zones indéniablement rattachées aux consonnes ... . La tenue ne suffit pas à l'oreille pour distinguer une consonne".

Or il nous semble que l'oreille est capable de distinguer au moins quelques consonnes comme le *f* ; nous remarquons toutefois que l'audition d'une consonne tronquée peut bien déformer cette consonne (exemple un *s* est entendu comme un *t*. En général, la brusque interruption produit l'occlusive correspondante.

Sur une idée de Benzécri J.P., nous partons d'une consonne (une tranche) et avec la transformée de Fourier directe puis inverse, nous étirons le signal dans le temps pour faire durer la tenue de la consonne qui est écoutée après étirement.

Les résultats obtenus sont plutôt encourageants.

Il semble que l'opinion des auteurs qui assimilent les consonnes à des phénomènes exclusivement transitoires, sans existence propre, indépendante des voyelles, soit fondée sur l'aspect des tracés du VOCODER où, aux bandes intenses caractéristiques des voyelles, succèdent des zones quasi blanches pour les consonnes, la transition étant seule visible. En analyse des correspondances, le niveau de puissance est éliminé, seul compte le profil ; de ce point de vue, la consonne, même si elle est associée à une faible puissance, garde un dessin spectral aussi net que celui des voyelles qui l'encadrent ; la transition des formes (sur laquelle se base notre segmentation) est seule considérée sans que la courbe de puissance interfère avec elle.

CHAPITRE G

RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

Nous ne savons pas comment l'homme perçoit, analyse et reconnaît un message parlé ; il est donc impossible de simuler sur un ordinateur le très complexe processus humain et il n'est, de toute façon, pas assuré qu'un système de reconnaissance automatique doive copier celui de l'homme.

Cependant, la plupart des systèmes actuels de reconnaissance automatique de parole continue, se distinguent souvent par plusieurs niveaux de traitement, les plus utilisés étant les suivants :

- un niveau acoustique : codage, analyse et représentation du signal vocal.
- un niveau phonétique : segmentation et reconnaissance des segments.
- un niveau lexical : recherche et identification des mots à partir de la reconnaissance de niveau précédent.
- un niveau syntaxique : prise en compte des contraintes propres au langage.
- un niveau sémantique : interprétation du signal reconnu.

En réalité, il s'agit plutôt d'un système "à comprendre" qu'à un système "à connaître" la parole.

Nos travaux sur la reconnaissance automatique de la parole continue (valables aussi pour des mots isolés) se limitent aux deux premiers niveaux. Nous pensons améliorer notre "reconnaissance" par l'utilisation d'autres sources d'information et d'autres niveaux de traitement.

.../...

G.1. APPRENTISSAGE

Les difficultés rencontrées en reconnaissance automatique de la parole peuvent provenir de trois sources différentes : variations inter-locuteurs (différence d'élocution entre les individus), variations intra-locuteur (variations au cours du temps de l'élocution d'une même personne) et variations contextuelles (influences mutuelles des sons les uns sur les autres) . D'ailleurs un projet vocal n'a pratiquement jamais, toutes les conditions précédentes étant réunies, une même réalisation sonore ; il est donc impératif d'analyser tous les projets sonores, de découvrir leurs caractéristiques et de les classer, chaque classe ayant une description acoustique, phonétique, linguistique ou autre.

L'acquisition et le sotckage dans la mémoire de l'ordinateur des descriptions relatives à chaque classe constituent ce qu'on appelle communément l'apprentissage car ces descriptions permettront, par la suite, de localiser et d'identifier les unités à reconnaître.

Selon les systèmes, l'apprentissage est à refaire pour chaque locuteur, chaque texte et chaque langage. Nous sommes encore bien loin des systèmes de reconnaissance universelle.

Pour découvrir et décrire les classes, nous procédons d'abord par une analyse factorielle des correspondances. Tout profil spectral -correspondant à une tranche de temps fixe- est caractérisé par ses coordonnées dans l'espace des premiers facteurs (six ou sept). Les proximités des spectres permettront de découvrir les classes.

La description par les facteurs présente les avantages suivants :

- c'est, avant tout, une réduction fidèle des données ; la reconstitution du signal à partir des facteurs le prouve.
- l'analyse factorielle, avec utilisation de l'enveloppe spectrale, donne des facteurs stables quant aux variations de texte et de temps. Nous étudierons ultérieurement les changements éventuels des facteurs avec le locuteur et le langage.

.../...

- l'analyse des correspondances permet de projeter,dans l'espace des facteurs,tout élément nouveau - supplémentaire - opération nécessaire à l'étape ultérieure qu'est la reconnaissance.

Nous utilisons cette description du profil spectral pour les deux étapes suivantes de l'apprentissage.

G.1.1. Nombre de classes

Nous avons déjà expliqué que nos analyses et notre segmentation conduisent à adopter le phone comme unité phonétique. En partant d'un découpage aveugle de la chaîne parlée, nous laissons à l'analyse des correspondances, sans aucun a priori, le soin de regrouper les sepctres et de former parfois, à l'encontre de toute attente, deux,voir plusieurs,classes pour un phonème ; l'explication linguistique -qu'on a souvent pas cherchée- n'était pas toujours évidente, bien que des discussions avec Benzécri J.P. et Gzell nous aient permis de comprendre et d'interpréter cette classification. Nous pensons que c'est un sujet qui mérite des réflexions plus profondes qui mèneront sans doute à des résultats linguistiques intéressants. Cependant, nous n'avons pas construit une nouvelle classe à chaque fois qu'un ou plusieurs spectres se détachent, "bizarrement",de leur groupe "parent" dans l'espace des facteurs ; les classes formées doivent être "significatives". Ainsi le nombre de classes obtenues par l'analyse factorielle se situe entre 44 et 50. Il est difficile de fixer d'une manière précise le nombre de classes : deux classes, dérivées de deux phonèmes différents se superposant presque l'une sur l'autre, peuvent constituer une seule classe comme par exemple la classe de la voyelle U et celle de la semi-voyelle W.

G.1.2. Caractérisation numérique d'une classe

Tout spectre -correspondant à une tranche de temps fixe du signal- est un point de l'espace des facteurs ; il est ainsi caractérisé pour chaque axe par trois valeurs numériques :

- l'abscisse du point : sa position sur l'axe
- sa contribution au facteur
- sa corrélation avec le facteur

Généralement les spectres d'une classe de phone ont, surtout pour les facteurs où ils s'expriment le mieux, des distributions de ces trois statistiques avec des faibles dispersions comme le montre le tableau F.1. relatif à la consonne B.

.../...

TITRE: ANALYSE 2

!	II	!	QLT	POID	INR	1=F	COR	CTR	2=F	COR	CTR	3=F	COR	CTR	4=F	COR	CTR	5=F	COR	CTR	6=F	COR	CTR	7=F	COR	CTR
187	BBBB	!	998	0	0	-129	32	0	-82	13	0	528	525	0	-179	61	0	1	0	0	-233	103	0	-362	248	1
188	BBBB	!	998	0	0	-132	28	0	-83	11	0	549	471	0	-186	55	0	0	0	0	-122	24	0	-490	376	1
189	BBBB	!	998	0	0	-130	27	0	-82	11	0	539	461	0	-188	57	0	0	0	0	-136	30	0	-489	380	1
297	BBBB	!	999	0	0	-129	28	0	-78	10	0	526	464	0	-170	49	0	0	0	0	-151	39	0	-469	369	1
298	BBBB	!	999	0	0	-129	30	0	-79	11	0	523	482	0	-173	54	0	1	0	0	-224	89	0	-406	291	1
299	BBBB	!	999	0	0	-128	31	0	-79	12	0	528	515	0	-177	58	0	1	0	0	-232	101	0	-377	264	1
300	BBBB	!	998	0	0	-128	39	0	-79	15	0	509	609	0	-150	54	0	1	0	0	-194	89	0	-272	175	0
314	BBBB	!	998	0	0	-130	29	0	-81	12	0	519	462	0	-161	45	0	0	0	0	-201	70	0	-456	357	1
315	BBBB	!	998	0	0	-132	28	0	-82	11	0	521	430	0	-155	39	0	0	0	0	-150	36	0	-511	415	1
316	BBBB	!	998	0	0	-131	28	0	-82	11	0	526	442	0	-163	43	0	0	0	0	-182	54	0	-486	380	1
317	BBBB	!	998	0	0	-131	26	0	-82	10	0	539	436	0	-175	46	0	0	0	0	-124	23	0	-529	420	1
1628	BBBB	!	999	0	0	-126	27	0	-80	11	0	542	489	0	-192	62	0	0	0	0	-137	32	0	-455	346	1
1629	BBBB	!	998	0	0	-124	23	0	-77	9	0	540	434	0	-186	52	0	-1	0	0	0	0	0	-551	453	1
!		!	3963289.		0			0			0			0			0			0			0			0	.12
BBBB		!			!	-129.1538		!	-80.4615		!	529.9231		!	-173.4615		!	.2308		!	-160.4615		!	-450.2308		!

Tableau G.1

Description par l'abscisse, la corrélation et la contribution aux sept premiers facteurs de 13 spectres de la consonne B.
Les valeurs de la dernière ligne sont les coordonnées du centre de gravité de la classe B dans l'espace des sept premiers facteurs.

Ainsi toute classe peut être localisée par une moyenne, voir un intervalle ; nous disons alors une fenêtre sur l'espace, de l'une ou de plusieurs de ces trois séries de valeurs. Cette procédure présente les avantages d'une faible mémorisation de l'apprentissage et d'un temps de calcul réduit.

G.2. RECONNAISSANCE AUTOMATIQUE

Nous partons de l'analyse d'un texte dont nous supposons connaître le contenu phonétique de chacun des spectres après l'avoir écouté plusieurs fois, observé le signal et comparé avec le texte. Nous disons que nous supposons connaître le contenu de chaque spectre car l'analyse et l'expérience que nous décrivons plus loin nous réservent parfois des surprises.

L'analyse factorielle permettant de dégager les classes de phones, nous prenons pour représentant de la classe son centre de gravité.

Pour reconnaître un segment, nous calculons sa distance aux centres de gravité des différentes classes et nous l'affectons à la classe la plus proche.

Deux procédures de reconnaissance sont alors possibles.

G.2.1. Reconnaissance par spectre de durée fixe

Cette première procédure tente la reconnaissance sans une préalable segmentation dans le souci d'éviter toutes les variations dûes au locuteur et au texte. En effet le signal étant découpé en spectres successifs de durée constante de 50 ms (ou de 25 ms), après avoir isolé les spectres silences (dont le total de l'énergie est inférieur à une valeur fixée d'avance), nous procédons à la reconnaissance spectre par spectre en affectant chaque spectre à la classe dont le centre de gravité est le plus proche.

Nous disons, tout de suite, que l'inconvénient majeur de cette procédure résulte de la présence des segments transitoires qui sont d'une grande variabilité et qu'il est pratiquement impossible de reconnaître dans l'un ou l'autre des phonèmes adjacents ; même si cette tentative n'est toujours pas vouée à l'échec. En plus, il est généralement difficile de délimiter la transition et de déterminer les spectres "purs" ; toutefois l'examen des 817 spectres de 25 ms permet de dégager 187 spectres

transitions, soit un pourcentage d'environ 23%, ce qui diminuerait forcement le taux de reconnaissance. Nous verrons comment nous avons essayé de surmonter cet inconvénient.

La comparaison des distances aux centres de gravité des classes de phones nous permet d'avoir plusieurs réponses possibles ordonnées par ordre décroissant de vraisemblance (ordre croissant des distances). Nous retenons pour chaque spectre cinq à six réponses classées en positions 0, 1, 2, 3, 4 et 5.

L'étape suivante consiste à exploiter, au mieux, cette chaîne de réponses multiples, en utilisant parfois des informations complémentaires.

G.2.1.1. Affectation du spectre à la classe la plus proche

Cette opération qui est la plus rapide, mais aussi la plus élémentaire, affecte le spectre au centre de gravité le plus proche - réponse qui est à la position zéro.

Pour chiffrer la réussite de reconnaissance, nous prenons le taux de reconnaissance des segments du texte qui a servi à l'analyse ou d'autres segments projetés comme éléments supplémentaires qui donnent un taux voisin de "bonnes" reconnaissances. Rappelons que les segments transitions (de l'ordre de 23% du nombre total des spectres) ne sont pas supprimés. Le pourcentage de "réussite" est de l'ordre de 42% des spectres. Nous exprimons nos réserves quant au vrai taux de reconnaissance qui est, en réalité, nettement supérieur à celui-ci comme nous l'expliquerons ultérieurement.

Cependant nous observons contrairement à un certain nombre de chercheurs que les consonnes sont beaucoup mieux reconnues que les voyelles. Ainsi les bonnes reconnaissances des consonnes varient entre 60 et 62% selon les analyses. Ceci s'explique peut-être par la grande variabilité des voyelles avec leur environnement phonétique, ce qui entraîne une plus grande dispersion des classes des voyelles qui restent tout de même regroupées dans des zones bien définies de l'espace des facteurs. L'étude des transitions a confirmé l'existence théorique de l'influence de la consonne initiale sur la voyelle suivante et aussi de la consonne finale sur la voyelle précédente.

Or en arabe, la voyelle *a* est la plus fréquente et souffre parmi les voyelles du plus faible taux de reconnaissance, ce qui conduit au tassement du taux de reconnaissance général.

Ajoutons que certains phones sont souvent correctement reconnus comme le *b*, d'autres le sont moins bien et que nous observons des confusions, ou mauvaises reconnaissances, de types bien connus comme entre *m* et *n* ; d'ailleurs, certains chercheurs ont tenté l'expérience de remplacer dans le signal d'un mot, le spectre du *m* par celui du *n* et ont trouvé que l'audition du signal modifié ne diffère guère de celle du signal original. Ajoutons que si nous tolérons certaines confusions (comme entre *d* et *d* emphatique, *n* et *m* ...), le taux de bonnes reconnaissances augmentera de 5 à 8% environ selon les analyses.

G.2.1.2. *Exploitation de la chaîne de réponses multiples*

Les classes des phones ne sont pas toujours disjointes dans l'espace des facteurs. Il existe donc des zones de recouvrement entre deux, voire plusieurs zones de classes , ce qui conduit, en bonne partie, à des erreurs de reconnaissance aux premières réponses possibles. En effet, nous avons constaté que le spectre est reconnu dans les quatre premières réponses avec un taux variant entre 74% et 80% selon les analyses. Il est donc intéressant de tirer profit de toutes les réponses, surtout les premières, pour améliorer la reconnaissance automatique; nous imaginons plusieurs méthodes.

G.2.1.2.1. Matrice de transition

La matrice de tansition est une matrice carrée (nxn, où n est le nombre de phones) où l'élément (i,j) représente le nombre de fois où le phone i précède (ou suit) le phone j - on peut en déduire les fréquences d'apparition du phone i avant (ou après) le phone j. Les éléments de cette matrice ne sont pas tous non nuls ce qui permet déjà de supprimer une ou plusieurs réponses de la chaîne fournie plus haut. Il est même possible d'utiliser les fréquences pour probabiliser les réponses et en tirer meilleur profit.

Nous pensons que l'utilisation de cette matrice pourrait être fort utile dans le cas d'un vocabulaire limité (même en parole continue) où les transitions possibles sont connues et limitées. Nous n'avons pas jugé opportun d'utiliser cette matrice avec l'hypothèse d'un vocabulaire non limité.

.../...

G.2.1.2.2. Chemin le plus proche de la position zéro

Pour chaque séquence de parole continue (séquence se trouvant par exemple entre deux silences car nous n'utilisons, jusqu'à maintenant, aucune segmentation phonémique) et pour chaque réponse de la chaîne, jusqu'à la disparition de la réponse d'une chaîne, nous construisons un indice dont la valeur est la somme des indices de position de la réponse. Ainsi une réponse, pour un nombre limité de segments successifs, suit un petit chemin qui s'approche ou s'éloigne de la position zéro. Nous attribuons à cette suite de segments la réponse qui a le chemin le plus proche de zéro. L'inconvénient de cette méthode est de fausser la reconnaissance des sons brefs comme un *t* entre deux *a* : ce *t* sera confondu avec les *a*. Il est possible de remédier à cet inconvénient ; car il est possible de distinguer une transition d'un son bref, ce dernier étant reconnaissable sur un ou plusieurs facteurs, et donc de corriger la mauvaise reconnaissance d'une transition sans nuire à la reconnaissance d'un son bref.

G.2.1.2.3. Indice de proximité

Nous avons imaginé la construction d'un indice de proximité pour chacune des cinq ou six réponses de la chaîne. Le calcul de cet indice s'inspire de celui du chemin le proche de zéro.

En effet, comme nous prenons, en général, des spectres décalés dans le temps, donc chevauchant, mis à part les sons brefs dont la reconnaissance ne pose pas de problèmes particuliers et peut être sauvegardée dans les opérations qui suivent, un phone occupe généralement plusieurs segments. Pour tirer profit des chaînes de réponses données aux segments qui précèdent et qui suivent le segment à reconnaître, nous prenons toutes les réponses se trouvant dans une pyramide basée sur la réponse zéro ; cette pyramide peut être réduite ou agrandie en faisant varier un paramètre qui représente le nombre maximum de pas horizontaux et (ou) verticaux possibles en se déplaçant de la position zéro aux autres réponses de la pyramide. Nous attribuons à chaque réponse (case de la pyramide) un indicateur de distance à zéro qui représente le nombre de pas horizontaux et (ou) verticaux entre les deux réponses.

Puis pour chacune des réponses 0, 1, 3, 4 ou 5 suivant la grandeur de la pyramide, nous calculons l'indice de proximité de la manière suivante :

.../...

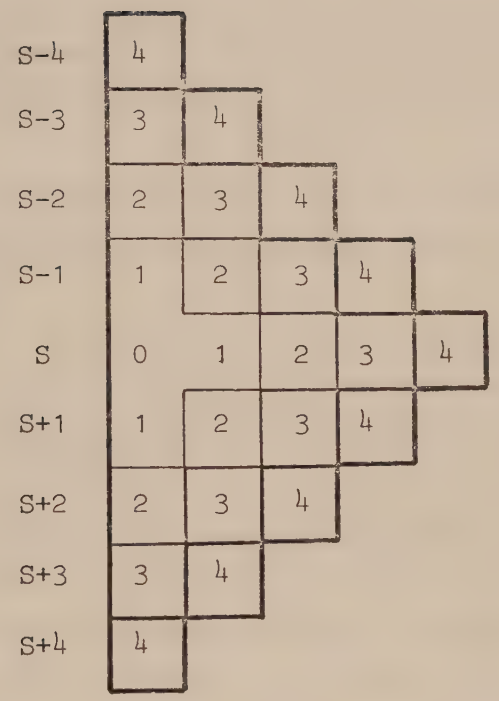


Tableau G.2

Prenons le cas du choix entre les trois premières réponses au spectre S (positions : 0, 1, 2), alors l'indice de la réponse zéro est égal à zéro plus 1, 2 ou 3 suivant que cette réponse se trouve dans la chaîne du segment S-1 aux positions respectives 1, 2 ou au-delà (pareil pour le segment S+1) plus 2 ou 3 suivant que cette réponse se trouve en première position ou ailleurs dans la chaîne du segment S-2 (pareil pour le segment S+2), voir tableau G.2.

Nous calculons d'une manière analogue les indices pour les autres réponses (1 et 2) au segment S. Les trois indices étant calculés, nous changeons le classement des réponses en permutant avec la réponse zéro celle qui a le plus petit indice.

Cette procédure permet de repêcher une réponse correcte se trouvant en seconde ou troisième position et parfois, mais très rarement, en quatrième position.

L'utilisation de cet indice de proximité n'est pas sans inconvénient pour certains segments mais dans l'ensemble elle permet d'améliorer le taux de reconnaissance de 5 à 8%.

Le nombre de segments qui entrent en jeu pour calculer cet indice dépend du débit de la parole et de la valeur du chevauchement des spectres.

G.2.1.2.4. Chaînes de réponses de longueur variable

Nous examinons les distances du spectre inconnu aux différents centres de gravité. Au lieu de prendre en considération toutes les réponses, nous pouvons les limiter en abandonnant les distances supérieures au premier saut "significatif" (saut à déterminer en fonction des données). Cela réduit les possibilités du choix (évite des erreurs) mais complique le calcul de l'indice de proximité : les chaînes de réponses n'ont plus les mêmes longueurs.

G.2.1.2.5. Utilisation de la corrélation et de la contribution

Nous avons dit qu'un spectre est caractérisé, pour chacun des facteurs, par son abscisse, sa corrélation et sa contribution. Bien que l'utilisation des coordonnées seules contribue grandement à une bonne première reconnaissance, il est intéressant de conjuguer les trois critères pour pouvoir trancher entre deux réponses successives possibles à un segment. En effet l'examen d'un certain nombre de confusions permet de les lever par comparaison de la corrélation et (ou) de la contribution du segment à reconnaître avec celles des réponses sur un ou plusieurs axes qui caractérisent ces réponses.

.../...

Nous nous efforcerons d'introduire ces critères dans notre programme de reconnaissance.

G.2.1.3. Augmentation du nombre de facteurs

Tous les résultats déjà cités sont obtenus avec une représentation du nuage des spectres dans l'espace des sept premiers facteurs. Nous avons déjà dit que l'inertie cumulée de ces facteurs et la reconstitution du signal nous permettent d'affirmer qu'ils représentent l'essentiel de l'information.

Pour confirmer ceci, nous avons essayé des reconnaissances des spectres dans les espaces des 8, 9, ... et 14 premiers facteurs. L'amélioration du taux de reconnaissance est très légère et n'est pas fonction croissante du nombre de facteurs utilisés. Cette amélioration est au meilleur des cas de l'ordre de 2 à 5% et curieusement c'est le phonème α dont la reconnaissance s'améliore mais pour d'autres phonèmes la reconnaissance est moins bonne qu'avec sept facteurs.

G.2.2. Reconnnaissance après segmentation

La reconnaissance par segment de durée fixe est intéressante à plusieurs égards surtout du fait de sa quasi-indépendance du débit de parole et peut être du locuteur. Cependant la reconnaissance des segments transitions demeure difficile. L'amélioration des résultats dépend d'une éventuelle détection des transitions -opération qui n'est pas aisée- pour les isoler et ne reconnaître que les segments restants du signal.

L'utilisation d'une segmentation préalable du signal va dans ce sens. En effet, une fois le segment déterminé et selon sa durée dans le temps -donc selon le nombre de spectres de durée fixe qui le composent- nous le détachons des spectres périphériques (un ou deux de chaque côté) et nous entreprenons sa reconnaissance.

Pour reconnaître un segment, ainsi déterminé, deux procédures sont possibles :

- reconnaître chacun des spectres -de durée fixe- qui le composent et déduire une réponse à partir de l'ensemble des chaînes de réponses données à ses spectres et ceci par l'une des méthodes décrites au G.2.1.2. .
- représenter numériquement ce segment par des moyennes des caractéristiques numériques de ses spectres (le centre de gravité par exemple) ; puis reconnaître ce nouvel élément.

Les deux principaux problèmes qui affectent la reconnaissance après segmentation sont les suivants :

- la segmentation -loin d'être parfaite- pose le problème de la reconnaissance d'un segment qui devait être découpé en deux (ou plusieurs) segments et qui, dans le meilleur des cas, peut être reconnu en l'un des sons qui le composent. Que devient alors le signal (mot ou phrase) "reconnu" ?. L'oreille (éventuellement une machine) peut-elle reconstituer le message ?. Les expériences sur les communications téléphoniques prouvent que l'homme, par un système complexe non connu, peut comprendre un message quand certaines de ses parties sont imprécises. Par une machine "à comprendre" pouvons-nous, avec un système très élaboré, atteindre ce résultat ?.

- la segmentation conduit à des segments de longueurs -durées- inégales, ce qui complique la tâche de la reconnaissance. Plusieurs systèmes de reconnaissance de la parole cherchent une normalisation temporelle des segments avant toute opération d'identification. Certaines normalisations sont faites par programmation dynamique.

G.2.3. Reconnaissance hiérarchique

Toutes les expériences concordent avec le constat suivant : les critères acoustiques ne sont pas ultimement suffisants pour une reconnaissance parfaite de la parole continue et le recours à d'autres sources d'informations paraît nécessaire. Les analyses du type syntaxico-sémentique permettent de trancher entre plusieurs réponses possibles, de reconsidérer une décision et par suite reprendre la reconnaissance voir même la segmentation dans une tranche du signal. Plusieurs niveaux de décisions hiérarchisées permettront d'améliorer la reconnaissance. Ce travail nécessite une durée plus importante que celle du signal : aucun des systèmes actuels de reconnaissance de la parole continue ne fonctionne en temps réel.

G.3. ERREURS OBJECTIVES EN RECONNAISSANCE AUTOMATIQUE

Dans tout système de reconnaissance automatique deux adaptations réciproques se présentent entre l'homme et la machine pour contribuer à une meilleure performance du système donc à une réduction des erreurs de reconnaissance. En fait,

.../...

dans les systèmes les plus performants on ne sait pas exactement qui des deux s'adapte le plus à l'autre. Et puisque nous parlons d'erreurs, pouvons-nous toujours parler d'erreurs réelles et objectives du système ?.

Que doit reconnaître le système ? . Ce qui est dit , nous l'affirmons, ce n'est pas le texte comme il aurait dû être lu ou comme nous croyons l'avoir lu. Entre le projet sonore et sa réalisation, les différences ne sont pas perceptibles par le locuteur. Nous pensons que le système doit reconnaître ce qui est dit et non le corriger.

Pour confirmer cette réalité, nous avons donné un enregistrement de notre texte à Madame Benzécri qui ne parle pas l'arabe mais à qui, cette langue n'est pas totalement étrangère. Après avoir écouté plusieurs fois l'enregistrement, elle en a fait une transcription phonétique. La comparaison de cette reconnaissance et celle de notre système prouve que plusieurs erreurs, présumées comme telles, de segmentation et de reconnaissance ne sont pas objectivement des erreurs mais correspondent plutôt à la réalisation acoustique du projet sonore ; ce qui permet d'affirmer que notre système est plus performant que nous ne le croyions au début.

CONCLUSION

Nous avons travaillé sur un sujet difficile et très complexe où plusieurs disciplines et sciences sont impliquées ; nos recherches auraient pu être portées vers une seule orientation limitée et plus approfondie de statistique, de phonétique, de linguistique, de segmentation ou de reconnaissance, sans oublier l'informatique qui nous permet de tout envisager.

Nous avons approché le problème de l'analyse et de la reconnaissance de la parole continue sans avoir achevé toutefois d'exploiter la voie ouverte devant nous mais où nous poursuivons avec d'autant plus d'application que la reconnaissance automatique de la parole apparaît comme une nouveauté essentielle pour l'Homme et la civilisation.

ANNEXES

ANNEXE 1

CONFIGURATION DES SYSTEMES INFORMATIQUES ET ANALOGIQUES

UNITE CENTRALE

Hewlett Packard 2100 A, 32 Kmots de 16 bits, 1 registre d'entrée, 2 registres accumulateurs, addition, multiplication en virgule flottante cablées, 2 canaux d'accès direct à la mémoire autorisant une vitesse de transfert maximale de 1020400 mots par seconde. 52 périphériques peuvent être reliés simultanément.

MEMOIRES DE MASSE

1) Une unité de disques HP 7900 A à têtes mobiles, qui comprend 2 disques, l'un fixe, l'autre amovible comportant chacun 200 pistes de 48 secteurs. La capacité totale de stockage est de 5 millions d'octets et la vitesse de transfert moyenne est de 312000 octets/seconde.

2) Un dérouleur de bande magnétique HP 7970 E à pistes compatible IBM. La densité d'enregistrement est de 1600 bits par inch et la vitesse de défilement de la bande 45 inchs par seconde (variations à court terme de cette vitesse inférieure ou égale à ± 3%). Temps d'arrêt et temps de départ : 8,333 ms. La vitesse de transfert maximale (pour des blocs de longueur infinie) est de 72000 octets par seconde.

CONVERTISSEURS

1) Numérique analogique : 2 canaux indépendants comprenant chacun 1 convertisseur Analog Devices à 12 bits DAC 12 QM. Le temps de conversion est inférieur ou égal à 1 µs. Les deux canaux peuvent être couplés pour donner une précision de 13 bits sur un seul canal.

2) Analogique numérique : 1 canal comprend 1 échantillonneur Analog Devices SHA-2 A (temps d'échantillonnage 500 nanosecondes) et un convertisseur Hybrid Systems ADC 591 A - 12 A de 12 bits de précision. Le temps de conversion minimal est de 3,5 µs.

.../...

AUTRES PERIPHERIQUES

- Une imprimante General Electric (Termi Net 340)
- Une console Hewlett Packard 2649 C servant de pupitre de commande
- Un traceur BENSON 1102
- Un lecteur de ruban perforé HP 2748 A lisant jusqu'à 500 octets par seconde.
- Un perforateur de ruban (HP 8100 A) perforant jusqu'à 500 octets par seconde.
- Une unité de visualisation (HP 1300 A) ayant une définition d'image de 1/256 en abscisse et en ordonnée.

L'ensemble du système informatique fonctionne sous DOS (Disc Operating System).

MATERIEL ANALOGIQUE SPECIFIQUE AU TRAITEMENT DU SIGNAL SONORE

- Horloge externe : C'est un générateur sinusoïdal variable de 1 Hz à 1 MHz (Philips 5160). La fréquence du signal est préalablement divisée par 10 (gamme utile : C, 1 Hz à 100 KHz) et le signal transformé en impulsions rectangulaires de 200 ns de largeur et de même période que la sinusoïde. Cette suite d'impulsions détermine la cadence d'échantillonnage.

- Filtres : 2 canaux indépendants comprenant chacun un filtre passe bas fixe construits suivant le schéma fourni par Dunbar (Mathews 1969, p. 30), ayant une bande passante à 1 dB de Ø à 7000 Hz et une atténuation supérieure à 50 dB à 10000 Hz.

1 filtre réglable de 0,1 Hz à 100 KHz à 2 voies indépendantes ayant chacune une atténuation de 24 dB par octabe (Schlumberger FAB 24).

MAGNETOPHONES

Revox A 77 - Rapport signal-bruit = 60 dB environ.

AMPLIFICATEUR

Revox A 78 : 2 40 watts (4 - 3 ohms).

MICROPHONE

Boyer électrodynamique M88N.

BAFFLES

2 enceintes Acoustic Research AR3.

ANNEXE 2

TEXTE ARABE ANALYSE

كان اهالي دير القمر على موعد مساء الاثنين مع فرقة فرنسية للرقص الشعبي قدمت اولى حفلاتها داخل مبنى السرايا التي احتلّ الصِبيةُ سطحها بينما احتشد الأهل في باحتها وألفوا شبه حلقة رقصت داخلها الفرقة. جلس البعض على الأرض بينما تسلق البعض الآخر الشبابيك لعدم وجود كراس لهم.
وصفق الجميع وصفروا طويلا خلال الساعتين اللتين استغرقهما العرض. المشهد الاول ضم عشر رقصات والثاني تسعاً. وهي فنون من صميم الريف.
امتازت الرقصات بالحيوية والحركة السريعة والرشاقة التي يتمتع بها شباب الفرقة وفتياتها. ويبلغ معدل اعمار هؤلاء ١٧ عاما. اما الموسيقى فاقتصرت على عزف من شباب الفرقة على نوعين من الآلات: الأول ويشبه المزمار والثاني وهو نوع من الطبول.
والثياب مستوحاة من صور ورسوم وزخرفات قديمة داخل المتاحف الفرنسية وهي تعود الى القرنين الثامن عشر والتاسع عشر نفذتها جدّات الراقصات وهي ثقيلة الوزن ومتعددة الكشاكش والطيات ترافقها قبعة من القش تربط الى الخصر وتغطي الرأس قبعة من القماش الابيض المنشى تلتصق بالرأس وتتدلى فوق الاذن حتى تصل الى الكتفين. اما الاحذية فرقيقة جدا وشبيهة بأحذية راقصات الباليه وهي تساعد على القفز.
وفي مركز دائرة استقبال الشباب التابعة للمجلس الوطني لانماء السياحة التقينا الفرقة واعضاءها وهي فرقة هواة. رئيسها ومؤسسها قال انه كان ينتمي وزوجته الى فرقة فولكلورية غير ان ادارتها وطريقة العمل داخلها لم تعجباه فألف فرقة ودخلها ولداه وزوجته. وتضم الفرقة حاليا ١٤ عضوا من حملة شهادات الرقص. وتشترك عادة في مهرجانات الرقص العالمية. وفي ايطاليا في اطار المهرجان الدولي للرقص الشعبي ربحت الجائزة الاولى امام الاسرائيليين والرومانيين. وعلى رغم اننا فرقة غير محترفة فقد اشتركنا في مهرجانات اكس ولهذا المهرجان اهميته الخاصة اذ يحضره العديد من رجال الفن العالميين. وقد نشترك السنة المقبلة في احياء الذكرى المئوية الثانية لاستقلال الولايات المتحدة.
اعضاء الفرقة معظمهم من طلاب الصفوف الثانوية باستثناء طالب جامعي واحد يحضر اجازة في العلوم الاقتصادية وفتاة ممرضة.
تحدث طالب ثانوي يرقص ويعزف منذ ثماني سنوات عن اسفار الفرقة: كانت اطول سفرة قمنا بها مرهقة بعض الشيء. وعلى رغم ذلك ادّينا حفلاتنا تحت خيمة ولم يؤثر المطر على عدد المشاهدين الذي كان يتزايد يوماً بعد يوم.
وفي العام الماضي لم يكن تجاوب الجمهور معنا مشجعا اول الامر الا انه ما لبث ان صفق لنا. وفي احدى المحطات اضعنا حقائبنا وفيها ثياب الرقص، ولم نجد احدا يتكلم الانكليزية الا ان المصادفة كانت كبيرة فالتقينا راهبة قامت بدور المترجم وساعدتنا في العثور على الحقائب.

BIBLIOGRAPHIE

1. B.B.N.
"Speech understanding system - Final report".
B.B.N. Report n° 3438 , Boston 1976.

2. BENZECRI J.P.
"Pratique de l'analyse des données" 1 et 2 Dunod Paris 1980.

3. BENZECRI J.P.
"Pratique de l'analyse des données, 3, linguistique et lexicologie"
Dunod Paris 1981.

4. BRIDLE J.S., SEDGWIEK N.G.
"A method for segmenting acoustics patterns with applications to automatic speech recognition".
I.E.E.E. ICASSP Harford, pp 656-659, 1977.

5. CARRE R., HATON J.P., LIENARD J.S.
"Reconnaissance de la parole ; état de la recherche et du développement".
Collection des synthèses de SESORI; éditée par l'IRIA 1979.

6. CANTINEAU J.
"Cours de phonétique arabe".
Librairie C. Klineksieck, Paris, 1960.

7. CHABREL S., CHARBONNEAU G.
"Un système d'analyse de sons par ordinateur fonctionnant en mode conversationnel'
Onde Electrique, Vol. 8-9, p 358, 1976.

8. CHARBONNEAU G.
"Réalisation d'un système intégré d'analyse et de synthèse de sons par ordinateur. Application de ce système à l'étude de l'attribut de hauteur sonore et de sa perception".
Thèse de Doctorat, Orsay, 1976.

9. COLLIER R.
"Physiological correlates of intonation patterns".
JASA, Vol. 58, n°1, July 1975.

10. COOLEY J.W., LEWIS P.A.W., WELCH P.D.
"Historical notes on the Fast Fourier Transform".
Proc. IEEE, Vol. 55, n°10, Octobre 1967 .

11. COOLEY J.W., TUKEY J.W.
"An algorithm for the machine calculation of complex Fourier series".
Math. Computations, Vol. 19, pp 297-301, 1965.

12. DAVIS K.H. et al.
"Automatic recognition of spoken digits"
JASA, Vol. 24, pp 637-642, 1952.

13. DE MORI R.
"Recent advances in automatic speech recognition".
4th Int. Joint Conf. on pattern recognition, Kyoto (Japon), pp 106-124, 1978.

14. DE MORI R., LAFACE P., PICOLLS E.
"Automatic detection and description of syllabic features in continuous speech".
I.E.E.E. Trans. on ASSP, Vol. 24, n°5, pp 365-379, 1976.

15. DENES P.
"The design and operation of the mechanical speech recognizer at University College, London".
J. Brit. IRE, Vol. 19, n°4, pp 219-229, 1959.

16. DE SAUSSURE F.
"Cours de linguistique générale".
Payot, Paris.

17. DREYFUS M.
"Fortran IV"
Ed. Dunod, Paris, 1969.

18. DREYFUS-GRAF J.
"Sonograph and sound mechanics"
JASA, Vol. 22, pp 731-739, 1950.

19. ERMAN L.D.
"A functional description of the Hearsay II speech understanding system"
I.E.E.E. ICASSP, pp 799-802, 1977.

20. FLEISCH H.
"Etudes de phonétique arabe".
Mél. Univ. St Joseph, XXVIII, 6.

21. FOURIER J.B.
"Théorie analytique de la chaleur"
Paris, 1822.

22. GRAILLOT
"Compression et reconstitution des données issues du vocoder".
Publications du CNET, Vol. I, 1972-1973.

23. GREENWOOD D.D.
Critical bandwidth and the frequency coordinates of the basilar membrane".
JASA, Vol. 13, p 1344-1356, 1961.

24. GRESSER J.Y., MERCIER G.
"Exemple de reconnaissance automatique de la parole".
Commutation et électronique, n°32, p 49, 1971.

25. HARRIS Z.S.
"Distributionnal structure"
in word, 10/2-3, pp 146-162, 1954.

HARRIS Z.S.
"Co-occurrence and transformation in linguistic structure"
in language, 33-3, pp 283-340, 1957.

26. HATON J.P.
"Contribution à l'analyse, la paramétrisation et la reconnaissance automatique de la parole".
Thèse de Doctorat d'Etat, Nancy, 1974.

27. HATON J.P.
"Reconnaissance analytique de la parole aux niveaux acoustique, morphologique, lexical et syntaxique".
R.A.I.R.O. Informatique, Vol. 10, n°9, pp 57-75, 1976.

28. HATON J.P., LIENARD J.S., MERCIER G.
"Compte-rendu d'une mission d'étude en URSS"
Note technique CEI/CSE/28 - CNET Lannion, 1975.

29. HELMOLTZ H.
"Die Lehre von der Tonempfin dungen als physiologische Grundlage für die Theorie der Musik".
Ed. Vieweg and Sohn, Braunschweig (5° edition) (1896).

HELMOLTZ H.
"On the sensations of tone"
Traduction anglais par A.J. Ellis . Dover Public , New-York, 1954.

30. KAISER J.F.
"System analysis by digital computer"
Kuo and Kaiser Eds., New-York, Wiley, 1966.

31. KLATT D.H.
"Review of the ARPA speech understanding project"
92ème Meeting de l'Acoustical Society of America, San Diego, 1976.

32. KOENIG W., DUNN N.K., LACEY L.Y.
"The sound spectrograph"
JASA, Vol. 18, pp 19-49, 1946.

33. LESSER V.R. et al.
"Organization of the Hearsay II speech understanding system".
I.E.E.E. Trans. on ASSP, Vol. ASSP 23, n°1, pp 11-23, 1975.

34. LI, HUGHES, HOUSE
"Speech reconstitued from spectra of reduced dimensionnality"
JASA, 53, p 329, 1973.

35. LITTLER T.S.
"The physics of the ear"
Pergamon Press Ltd, London, 1965.

36. LAWRENCE B.T.
"The harpy speech recognition system"
Ph-D Thesis Carnegie Mellon Universy, Pittsburg, 1976.

37. MAISSIS A.
"Le traitement de l'information acoustique, étape fondamentale de la reconnaissance de la parole".
Thèse de Doctorat de 3ème cycle, Paris VI, 1973.

38. MARCAIS W.
"Le dialecte arabe des ūlâd Brahim de Saīda"
Paris, 1908.

39. MARKEL J.D., GRAY A.H.
"Linear prediction of speech"
Springer Verlag, 1976.

40. MARTINET A.
"La linguistique synchronique"
P.U.F. Paris, 1965.

41. MATTHEWS M.V.
"The technology of computer music"
M.I.T. Press, Cambridge, Mass., 1969.

42. MJASNIKOV M.I.
"Ob" ektivnoe rasposnavanie zvukov reci".
Zurnal Tehn. Fiz., T13, n°3, pp 109-115, 1943.

43. MOLES A.
"La structure physique du signal sonore"
Thèse Bibliothèque de la Sorbonne, Paris, 1952.

44. MONTEIL V.
"L'arabe moderne".
Librairie C. Klincksieck, Paris, 1960.

45. MOUSSA T.
"Problème de stabilité dans le calcul des facteurs et la discrimination en analyse des correspondances".
Thèse de Doctorat de 3ème cycle, Paris VI, 1972.

46. PIERCE J.R.
"Wither speech recognition"
JASA, Vol. 46, n°4, pp 1049-1051, 1969.

47. PLOMP R.
"The ear as a frequency analyzer".
JASA, Vol. 36, pp 1628-1636 ,1964.

48. POLO, VAN DEN KAMP, PLOMB
"Perceptual and Physical space of vowel sounds".
JASA, Vol. 46, pp 458-467, 1969.

POLO
"Analysis and synthesis of speech using a broad-band spectral representation".
Symposium on Leningrad, 1973.

49. QUERRE M., MERCIER G., GRESSER J.Y.
"Reconnaissance automatique de la parole : application de la programmation dynamique à l'identification des mots isolés. Résultats comparés sur 1000 mots".
Note technique CEI/CSI/44, CNET Lannion, 1974.

50. QUINTON P.
"Contribution à la reconnnaissance de la parole. Utilisation des méthodes heuristiques pour la reconnaissance des phrases".
Thèse de Doctorat d'Etat, Rennes, 1980.

51. RABINER L.R., GOLD B.
"Theory and applications of digital signal processing".
Prentice Hall, Inc., Englewood Cliffs, New Jersey, p 88 ,1975 .

52. REDDY R., ERMAN L.D., NECLY R.B.
"A model and a system for machine recognition of speech".
I.E.E.E. Trans. on ASSP, Vol. AU 21, N°3, pp 229-237, 1973.

53. RITEA H.B.
"A voice controllet data management system".
I.E.E.E. Symposium on speech recognition, Pittsburg, pp 28-31, 1974.

54. RODET X.
"Analyse du signal vocal dans sa présentation amplitude-temps. Synthèse de la parole par règles".
Thèse de Doctorat d'Etat, Paris VI, 1977.

55. RUNGE C.
"Zeit für Math und Physik"
Vol. 48, p 443, 1903 et ibid, Vol. 53, p 117, 1905 .

56. SAKAY T., DOSHITA S.
"The automatic speech recognition system for conversational sound".
I.E.E.E. Trans. on Electronic Computers, 1963.

57. SCHWARTZ L.
"L'intégrale de Fourier - Méthodes Mathématiques de la Physique".
Vol. 5, pp 1-34 Centre de Documentation Universitaire, Paris, 1963.

58. SHANNON C.E., WEAVER W.
"The mathematical theory of communication".
University of Illinois Press, 1962.

59. SHIGENAGA M., SEKIGUCHI Y.
"Speech recognition of spoken Fortran programs".
4th Int. Joint Conf. on Pattern Recognition, Kyoto, pp 990-994, 1978.

60. SHIKANO K., KOHDA M.
"A linguistic processor for train seat reservation in a conversational speech recognition system".
4th Int. Joint Conf. on pattern recognition, Kyoto, pp 1039-1041, 1978.

61. STEVENS K.M., KLATT D.H.
"Role of formant transitions in the voiced-voicelles distinction for shop".
JASA, Vol. 55, 3, 1974.

62. STRAKA G.
"Album phonétique"
Les Presses de l'université Laval.

63. TOMICHE N.
"Le parler arabe du Caire".
Mouton.

64. TUBACH J.P.
"Reconnaissance automatique de la parole".
Thèse de Doctorat d'Etat, Grenoble, 1970.

65. VICENS P.
"Aspects of speech recognition by computer".
Ph. D. Thesis, Stanford University, 1969.

66. VINTSINK T.K.
"Generative grammars and dynamic programming in speech recognition with learning"
I.E.E.E. ICASSP, Philadelphie, pp 446-449, 1976.

67. VINTSINK T.K. et al.
"Phoneme by phoneme of speech composed of words of a given vocabulary".
I.E.E.E. ICASSP, Philadelphie, pp 450-452, 1976.

68. WHITTAKER J.M.
"Interpolatory function theory".
Cambridge University Press, London, 1915.

TABLE DES MATIERES

INTRODUCTION

CHAPITRE A : LA PAROLE

A.1. L'APPAREIL PHONATOIRE

A.2. LA PHONATION

A.2.1. L'appareil respiratoire

A.2.2. Les cordes vocales et la glotte

A.2.3. Le conduit vocal

A.2.3.1. Excitation du conduit vocal
A.2.3.2. Transmission du conduit vocal

A.3. LA CLASSIFICATION PHONETIQUE

A.3.1. Les consonnes

A.3.2. Les voyelles

A.3.3. Les semi-voyelles

A.3.4. Les diphtongues

A.4. AUDITION ET PERCEPTION DE LA PAROLE

CHAPITRE B : LES RECHERCHES SUR LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

B.1. APPROCHES DE LA RECONNAISSANCE DE LA PAROLE

B.1.1. Approche globale

B.1.2. Approche analytique

B.2. RAPPEL HISTORIQUE

B.2.1. Les débuts

B.2.2. La reconnaissance des mots isolés

B.2.3. La reconnaissance de la parole continue

B.2.4. Les principaux systèmes de reconnaissance

CHAPITRE C : TRAITEMENT NUMERIQUE DU SIGNAL ACOUSTIQUE POUR UNE ANALYSE FACTORIELLE DE LA PAROLE

C.1. NUMERISATION DU SIGNAL

C.1.1. Choix du langage et du texte

C.1.2. Enregistrement analogique

C.1.3. Echantillonnage de la parole

C.1.3.1. Conditions théoriques
C.1.3.2. Conversion analogique numérique

C.2. ANALYSE SPECTRALE DU SIGNAL

C.2.1. Le logiciel

C.2.1.1. Structure générale du programme ASPECT
C.2.1.2. Les commandes

C.2.2. Le calcul des spectres de Fourier

C.2.2.1. Intérêt de l'analyse spectrale
C.2.2.2. Les formules et leur calcul
C.2.2.3. Choix des variables
C.2.2.4. Spectres successifs et spectres décalés

C.3. CODAGE DES SPECTRES POUR L'ANALYSE FACTORIELLE

C.3.1. Principe d'application de l'analyse des correspondances

C.3.2. Découpage en canaux des spectres et calcul des poids

C.3.3. Les individus (spectres)

C.3.4. Temps de traitement

CHAPITRE D : ANALYSE FACTORIELLE DES CORRESPONDANCES

D.1. CORRESPONDANCE ET TABLEAU DES DONNEES

D.2. NUAGES ET ESPACES

D.3. PROXIMITES ; EQUIVALENCE ET DISTANCE DISTRIBUTIONNELLES

D.4. INERTIE ET VARIANCE

D.5. AXES PRINCIPAUX D'INERTIE

D.6. PROPRIETES DES FACTEURS

D.7. CONTRIBUTIONS

D.8. FACTEUR SUR LES DEUX ENSEMBLES - FORMULES DE TRANSITION ET DE RECONSTITUTION

D.9. ELEMENTS SUPPLEMENTAIRES

CHAPITRE E : ANALYSE DE LA PAROLE

E.1. REDONDANCE

E.2. PARAMETRISATION ET CODAGE

- E.2.1. Méthodes temporelles
- E.2.2. Méthodes spectrales
- E.2.3. Codage spatial
- E.2.4. Statistique sans probabilité
- E.2.5. Réduction des données

E.3. ANALYSE DES TABLEAUX K(I,J)

- E.3.1. Influence de la durée de l'échantillon de parole
- E.3.2. Stabilité des résultats en fonction de l'échantillon de parole.
- E.3.3. Suppression du silence
- E.3.4. Choix de l'empiètement des canaux
- E.3.5. Sous-canaux supplémentaires
- E.3.6. Analyses : résultats et comparaisons
  - *E.3.6.1. Les facteurs stables et leurs inerties*
  - *E.3.6.2. Caractérisation du signal par zones de fréquence.*
  - *E.3.6.3. Avantages de l'enveloppe spectrale.*
  - *E.3.6.4. Interprétation des facteurs*
- E.3.7. Analyses de sous-ensembles de sons
  - *E.3.7.1. Analyse du signal sans X, S, Ø, J*
  - *E.3.7.2. Analyse du signal sans transitions*
  - *E.3.7.3. Analyse des consonnes*
  - *E.3.7.4. Analyse des sous-ensembles de consonnes*
  - *E.3.7.5. Analyse des voyelles*

E.4. RECONSTITUTION DU SIGNAL DE PAROLE

CHAPITRE F : SEGMENTATION DE LA PAROLE

F.1. LE PROBLEME DES UNITES

- F.1.1. La double articulation et l'écriture alphabétique
- F.1.2. Le trait pertinent et l'école de Prague

F.2. SEGMENTATION AUTOMATIQUE DE LA PAROLE

- F.2.1. Segmentation par analyse des correspondances
  - *F.2.1.1. Découpage aveugle*
  - *F.2.1.2. Segmentation*
- F.2.2. Particularité du dernier phonème du mot
- F.2.3. Influence du débit de parole et de l'articulation sur la segmentation

F.3. LES CONSONNES ET LEUR EXISTENCE PROPRE

CHAPITRE G : RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

G.1. APPRENTISSAGE

G.1.1. Nombre de classes

G.1.2. Caractérisation numérique d'une classe

G.2. RECONNAISSANCE AUTOMATIQUE

G.2.1. Reconnaissance par spectre de durée fixe

G.2.1.1. Affectation du spectre à la classe la plus proche

G.2.1.2. Exploitation de la chaîne de réponses multiples

G.2.1.2.1. Matrice de transition

G.2.1.2.2. Chemin le plus proche de la position zéro

G.2.1.2.3. Indice de proximité

G.2.1.2.4. Chaînes de réponses de longueur variable

G.2.1.2.5. Utilisation de la corrélation et de la contribution

G.2.1.3. Augmentation du nombre de facteurs

G.2.2. Reconnaissance après segmentation

G.2.3. Reconnaissance hiérarchique

G.3. ERREURS OBJECTIVES EN RECONNAISSANCE AUTOMATIQUE

CONCLUSION

ANNEXES

BIBLIOGRAPHIE